UNIVERSITAT DE BARCELONA

Structure and Traffic on Complex Networks

Jordi Duch i Gavaldà

Structure and Traffic on Complex Networks

STRUCTURE AND TRAFFIC ON COMPLEX NETWORKS

Memòria presentada per optar al títol de
Doctor per la Universitat de Barcelona

JORDI DUCH I GAVALDÀ
Dep. de Física Fonamental
Facultat de Físiques
Universitat de Barcelona
Martí i Franquès, 1
08028 Barcelona

Universitat de Barcelona
Barcelona, 2008

Els signants

FEM CONSTAR

Que el present treball que porta per títol

Structure and Traffic on Complex Networks

i que presenta en Jordi Duch i Gavaldà per optar al grau de Doctor per la Universitat de Barcelona, ha estat realitzat sota la nostra supervisió i que tots els resultats presentats i la seva anàlisi són fruit de la investigació realitzada per l'esmentat doctorand.

I perquè se'n prengui coneixement i tingui els efectes que correspongui, signem aquest certificat.

Director	**Tutor**
Alex Arenas Moreno	**Albert Díaz Guilera**
Professor Titular	Professor Titular
Dep. d'Enginyeria	Dep. de Física Fonamental
Informàtica i Matemàtiques	Universitat de Barcelona
Universitat Rovira i Virgili	

Barcelona, 16 de febrer del 2008

A la Mercè

Acknowledgments

Quan et llegeixes els agraïments de qualsevol tesi hi notes un aire distès i relaxat, molt allunyat del rigor i la serietat que es mostra en la resta del treball. No es estrany, doncs un cop comences a escriure aquest punt et dones compte que per fi ha arribat l'hora de tancar un cicle i passar pàgina. Aquesta secció proporciona un entorn ideal per a aturar-se un moment a reflexionar i recordar-se tota aquella gent que t'ha fet costat, directament o indirecta, durant el llarg camí que suposa fer un doctorat (en la majoria de casos sense comprendre massa bé el que fas). Espero no deixar-me ningú, tot i que si ho faig no es amb mala intenció.

Crec que es lògic quan penso en agraïments que la primera persona que em vingui al cap sigui l' Alex. Quan em va començar a parlar de sistemes complexos i em va abduïr cap al mon de la física em va deixar una mica "espantat"; però el temps m'ha demostrat que aquesta fe dipositada en ell s'ha acabat convertit en una de les millors decisions que he pres al llarg de la vida. El recull de la feina presentada en aquestes pàgines és tan sols una petita mostra del que he après al seu costat, però el que realment m'ha ensenyat va molt mes enllà del que es pot explicar amb simples paraules. Voldria extendre aquests mateixos agraïments a l'Albert, qui també ha estat ajudant-me en aquesta transició des de bon principi. Ha estat tota una aventura poder compartir articles, viatges i congressos amb vosaltres.

Durant aquests quatre anys com a becari de la URV hi ha hagut molta gent amb qui he compartit moltíssimes hores, he establert noves amistats i a qui haig d'agrair el seu suport. En primer lloc haig de destacar als companys de feina i de despatx, el Leon, el Sergi i l'Albert. Ha estat una gran experiència conviure amb vosaltres, tan a nivell personal com professional. Molta sort en les vostres carreres científiques, no tinc cap dubte que les coses us aniran molt bé. Amb el Juan i el Sergio no hem pogut treballar gaire temps plegats ja que es van afegir més tard al grup, però han estat sempre disponibles per a resoldre qualsevol dubte. Tambe voldria agraïr tots els altres companys del

Departament d'Enginyeria Informàtica i Matemàtiques, i en especial als qui m'han ajudat en moment puntuals o hem compartit classes plegats: el Pedro, el Carles, l'Helio, la Maria, i sobretot al Robert, amb qui em sembla que se m'esta acumulant el número de favors que li dec.

I ara que ja ha començat una nova etapa, voldria aprofitar per agraïr el suport de la gent de la Northwestern i de Chicago. Sobretot el Roger i la Marta que m'han ajudat en tot el que he necessitat des de que vaig arribar. I also want to thank all the people of the Amaral group. I've only been here for a few months but I'm sure that I couldn't find a group of friends like these anywhere else. Thanks for being there in a really hard time. And above all, my most sincere thanks to Luis for giving me the opportunity to enter the 'major leagues' of science.

M'agradaria recalcar que una de les coses que més m'ha sorprès de la vida científica es la seva 'vessant social'. Una de les grans avantatges d'aquesta vida es que et permet conèixer molta gent, gent que et vas trobant periòdicament, amb la comences discutint sobre qualsevol tema i amb la que al final acabes desenvolupant molt bones amistats. Des d'aquí voldria recordar totes les llargues hores de conferències que hem sofert plegats amb la Mariangels, el Marian, el Conrad, el Jesús, el Santo, i tots aquells amb qui he compartit alguna xerradeta i alguna que altra cerveseta. I també al Renato per donar-me l'oportunitat de poder fer una estada a Roma.

Per altra banda, potser on realment aprecio el suport rebut és quan penso en tota aquella altra gent que tens al teu costat fora de la feina, amb qui comparteixes esmorzars al Delta, sopars al xino, partits de futbol, estones a Pachito, o fent una partida a l'ordinador fins que surt el sol... Ja sabeu que us voldria posar a tots, però la llista es tan gran que no cabríeu!! Moltes gràcies per estar sempre allí, sobretot quan les coses estaven una mica complicades. En particular, m'agradaria donar-li especialment les gràcies als que m'heu ajudat en aspectes puntuals de la tesi: al Juando per totes les hores que va perdre donant-me suport i per convertir-se en la meva segona consciència, al Lluís per solucionar-me tots els problemes tècnics que m'han sorgit, i al Toni per ajudar-me amb el disseny de la tesi. També agrair-li al Marc Badia tot el suport logístic que m'ha donat en aquesta anys, qui ho havia de dir que al final acabaria abans que tu? Ànims que ja falta poc!

També voldria agrair de tot cor el suport que m'ha donat els darrers anys el que s'ha convertit en la meva segona família, i que m'han acollit com si fos un fill més. Sembla ser que tots aquells dissabtes que no em quedava a fer la partideta han servit per fer alguna cosa de profit!

Per acabar vull dedicar la tesi a tota la família, com a mostra de gratitud per ensenyar-me dia a dia aquelles coses que no s'expliquen a la universitat. A les padrines Maria i Carme que estaran preocupades perquè hagi marxat a treballar tan lluny, i als padrins que, encara que no estiguin amb nosaltres, segur que també haurien estat molt contents. A l'Ignasi i la Mireia, espero que seguiu treballant tant com fins ara, tots dos sou un bon exemple a seguir; sembla que seré el primer doctor de la família, però espero no ser l'únic. I sobretot vull agraïr als meus pares per donar-me sempre tot el que he necessitat, i fer-me costat en les decisions que he près. A tu papa per a estar cada dia darrere meu fent que sempre m'esforci per millorar. I sobretot, aquesta tesi te la dedico a tu mama. Si hi ha alguna cosa que m'ha donat forces durant aquests darrers quatre anys ha estat el teu exemple. Ens has ensenyat com la voluntat fa que puguis tirar endavant qualsevol cosa, fins i tot aquelles on ningú més hi creu. Espero que allà on estiguis et puguis sentir orgullosa de mi.

I només em falta agraïr a una persona, la que sempre ha estat al meu costat, compartint tots els bons moments i donant-me esperances en els dies mes difícils. Moltes gràcies Berenice, sense tu res d'això hauria estat possible.

Contents

Acknowledgments ix

List of Figures xvii

List of Tables xxviii

1. COMPLEX NETWORKS: BETWEEN ORDER AND RANDOMNESS 1

  1 Describing complex networks 3

    1.1 Statistical properties of complex networks 5

    1.2 Network models 9

  2 Community structure of complex networks 16

  3 Traffic dynamics 18

    3.1 Traffic dynamics on complex networks 19

  4 The Internet viewed as a complex system 20

    4.1 Discovering Internet Topology at the AS level 21

    4.2 Modeling the Internet 23

    4.3 Internet Traffic modeling 25

  5 Scope of the work 26

2. DETECTING COMMUNITY STRUCTURE IN COMPLEX NETWORKS 29

  1 Defining the community structure 30

  2 Detecting community structures 31

    2.1 Link removal methods 34

    2.2 Agglomerative methods 36

    2.3 Methods based on maximizing modularity 38

    2.4 Spectral analysis methods 39

    2.5 Other methods 41

  3 Comparative evaluation 44

    3.1 Accuracy of the methods 44

    3.2 Efficiency of the methods 48

3.3 Which algorithm should we use? 49
4 Summary 50

3. DETECTING COMMUNITY STRUCTURE USING EXTREMAL OPTIMIZATION 53
1 Extremal Optimization Algorithm 54
1.1 Optimizing the modularity 56
1.2 Implementation details 60
1.3 Testing the EO algorithm performance 61
2 Detecting weighted and directed communities 62
3 Increasing the efficiency and accuracy of community detection algorithms 65
3.1 Improving recursive algorithms 66
3.2 Size reduction of the network 68
4 Uncovering the community structure of real networks 74
4.1 Community detection benchmark 76
5 Summary 78

4. SCALING OF FLUCTUATIONS IN TRAFFIC ON COMPLEX NETWORKS 81
1 Scaling of fluctuations on complex networks 82
2 A simple traffic model 83
3 Effect of the dynamical parameters on the scaling exponent 84
3.1 Effect of the time window 85
3.2 Effect of the traffic variability 86
3.3 Effect of the congestion 88
4 Scaling exponent of Internet traffic 89
5 Summary 91

5. DYNAMICAL ROBUSTNESS OF A COMMUNICATION PROCESS 93
1 Robustness of Complex networks 94
2 Determining the Robustness of a Communication Process 95
2.1 Topological robustness 95
2.2 Dynamical robustness 97
2.3 How to determine the dynamical robustness? 98
3 Effects of a node removal on the onset of congestion 99
3.1 ER and SF networks 99
3.2 Regular lattice 101
3.3 Effect of the knowledge radius 102
4 Comparing topological and dynamical robustness 106
5 Summary 107

6. CONCLUSIONS AND PERSPECTIVES 109
1 Conclusions 109
2 Perspectives 112

Appendices 115
Evolution of the Internet Autonomous System Topology 115
1 Mapping the Autonomous System topology 116
2 Analyzing the evolution of the AS maps 118
3 Summary 124
Relationship between directed and undirected modularities 125
Analytic network reduction preserving modularity 127
1 Size reduction preserving modularity 127
2 Reductions for undirected networks 128
3 Reductions for directed networks 130
Resum de la Tesi 133
Publication list 145

List of Figures

1.1 Evolution of the number of hosts connected to Internet since the creation of the network in 1969 in a logarithmic scale. Data has been reproduced from the Hobbes' Internet Timeline, which is available for download at http://www.zakon.org/robert/internet/timeline/ 2

1.2 a) Example of an unweighted and undirected graph composed by nodes (points) and links (lines). b) Representation of the graph using an adjacency matrix, where the value of $a_{i,j}$ is 1 if there is a link between nodes i and j and 0 otherwise. c) Implementation of the graph using linked lists. The first column are the nodes, and the list that follows each nodes contains the destination of the links that each node has. 5

1.3 Left: Example of a 1D lattice, represented as a ring with periodic boundaries, where each node is connected to neighbors at distance 1 and 2. Right: Example of a 2D lattice in a rectangular grid that also presents periodic boundary conditions. 10

1.4 Left: Example of a random network generated with the Erdös-Rényi method, with $N = 30$ and $p = 0.1$. Right: Degree distribution of an ER network with $N = 10000$ and $p = 0.02$, with a mean degree $\langle k \rangle \sim 200$. 11

1.5 Left: Random rewiring procedure introduced by Watts and Strogatz to create a network with high clustering coefficient and low average path length. Right: Normalized average path length L and clustering coefficient C as a function of the rewiring probability p. The shadowed area correspond to a range of values of p that have the small-world property, high clustering and low average path length. 14

1.6 Left: Example of a network without a characteristic scale. In this type of networks we usually observe a few high degree nodes that act as hubs, and a large number of peripheral nodes with a low number of connections. Right: Degree distribution of a network generated using the preferential attachment method of Barabási and Albert. The exponent of the power law is ~ 2.8. The average degree of the network is ~ 6 and the maximum degree is ~ 1600. 16

1.7 Example of the community structure in a complex network. The nodes can be classified into groups where the number of internal links (links between nodes of the same group) is higher than the number of external links (links with the rest of the graph). 17

1.8 The figures present three different views of the Internet topology generated by three different projects. Top: Geographical distribution of the ASs, represented as an arc map, see http://mappa.mundi.net/maps/maps_008/ for more information. The picture has been created by visualization researchers at Bell Laboratories-Lucent Technologies, ©Stephen Eick, Bell Labs. Bottom left: Detailed map of the Internet Router level from the Opte Project http://www.opte.org/. Used under the Creative Commons License. Bottom right: Hierarchical structure of the Internet AS level introduced in (Carmi et al., 2007). The size of the nodes represent their degree and the color their position in the nodes hierarchy. 23

1.9 Example of an AS Internet topology map generated using the Transit-Stub model. This topology generator creates an Autonomous Systems structure using the two main hierarchical elements of the Internet, the transit and the stub domains. 24

2.1 Example of the limited resolution of the modularity introduced in (Fortunato and Barthélemy, 2007). The network is composed by identical cliques (complete graphs with m nodes) connected by single links. The methods that optimize the modularity will group the cliques (dotted lines) if there are more than $\sqrt{L}$ cliques, instead of detecting the smaller but highly dense connected groups. 33

2.2 Shortest path centrality (betweenness) is the number of shortest paths that go through a link or node. In this simple case, the link with the largest link centrality is that joining nodes 4 and 5. 34

2.3 Resistor networks and current flow centrality. The links in the network are considered as unit resistances. By choosing a pair of nodes to be a source of unit voltage s and sink t, one can can calculate the current flow through any link using Kirchoff's laws. Summing this value for every pair of nodes gives the total current flow betweenness of a link. In this case the biggest current flow is through link joining nodes 4 and 5. 35

2.4 a) Overlapping communities around a given node. We can observe that one node can belong to more than community at the same time. The communities can overlap and share more than one node between them b) An example of overlapping k-clique communities at with $k = 4$. The red nodes belong to more than one community. For more information about the method see (Palla et al., 2005) 37

2.5 a) Components of the first non-trivial eigenvector for a *ad hoc* network with 4 communities (see Sec. 3). b) All communities can be clearly identified when the components of more than one eigenvector are used as coordinates in M-dimensional space where M is the number of eigenvectors used. Here $M = 2$. 40

2.6 How clustering is related to curvature according to (Eckmann and Moses, 2002). For a node i, the shortest path distance between any of its neighbors will be either 1, if the neighbors are linked, or 2, if they are not. The average distance between the neighbors can give a measure of curvature. Positive curvature is depicted in (a) and negative curvature is depicted in (b). Both triangles have sides of length unity, and the angle between the two is the same, but the distances are different. 41

2.7 The q-Potts model as applied to a small network with communities. Each node is assigned one of q spins. As the Hamiltonian of the system is minimized, the spins in a tightly connected community take equal values, which are different to those of spins located in other communities. 43

2.8 Algorithm sensitivity as applied to ad hoc networks with $n = 128$, the network divided into four communities of 32 nodes each and total average degree z_{out} fixed to 16. For low z_{out}/k the communities are easily distinguished. For higher z_{out}/k this becomes more complicated. Both measures of comparing original communities to ones found by the detection method are shown. The normalized mutual information measure is more discriminatory and appears more sensitive to errors in the community identification procedure. The results are shown for Newman's fast algorithm (Newman, 2004b). 45

2.9 Comparing algorithm sensitivity using ad hoc networks with predetermined community structure. The x-axis is the proportion of connections to outside communities z_{out}/k and the y-axis is the fraction of nodes correctly identified by the method measure as described in (Newman, 2004b). The labels here correspond to the different methods and are listed in table 2.1. 47

2.10 The fraction of correctly identified nodes at three specific values of z_{out}, 6, 7 and 8 for all available methods and for networks with fixed $k = 16$. Note that for the FLM method, the data for $z_{out} = 8$ were not available. Here we can see that most of the methods are very good at finding the "correct" community structure for values of z_{out} up to 6. At $z_{out} = 7$ some methods begin to falter but most still identify more than half of the nodes correctly. At $z_{out} = 8$, when on average half the links are external, the SA method is still able to identify over 80 % of the nodes correctly. 48

2.11 Left: Example of a computer generated network with communities of different sizes. In this case we can control the size, the internal cohesion and the external cohesion of each community, see (Danon et al., 2006) for more details. Right: Example of a computer generated network with hierarchical community structure. Each node has a high probability to create a link to nodes of the same internal cluster (z_{in1}), a lower probability to be linked to nodes of the external cluster (z_{in2}) and almost no probability to be linked with nodes of the rest of the network (z_{out}). See (Arenas et al., 2006b) for more details. 49

3.1 Example of the simulated annealing heuristic search process in a one-dimensional configuration space. Each configuration has associated a given energy value. With the simulated annealing we look for the configuration that minimizes the energy of the system. The red ball represents one run of the simulated annealing that is trapped in a local minima. When we increase the temperature of the system, we allow the ball to jump between local minima, potentially making any configuration accessible. 55

3.2 Left: Random initialization of the Zachary network into two partitions, red and green. Right: Here we identify five different communities by looking at the connected components in each partition. Each color defines a different community. 58

3.3 Top: Network after edge removal at each recursive cut. Bottom: Evolution of the Q value in the at each step of the adaptation process. Separation bars indicate recursive divisions of the graph performed at maximum Q. 59

3.4 Fraction of nodes classified in the same partition over 100 realizations of the algorithm. The color of the position (i,j) corresponds to the fraction of times that nodes i and j belong to the same partition. 60

3.5 Fraction of nodes correctly classified using computer-generated graphs described in text. Each point is an average over 100 different networks. Inset: Average of the maximum modularity obtained in each case. 62

3.6 Fraction of nodes correctly classified using computer-generated graphs with directed and weighted links. We fix the number of links, the weight of the external links to 1, and we increase the weight of the internal links. Dotted lines represent the accuracy of the detection process when we remove the direction of the links. 65

3.7 Left: Hierarchical tree that represents the recursive process behind the EO algorithm. The final partitions are defined by the leaves of the tree. Right: Combining the EO with Newman's fast algorithm we can explore other possible configurations that can still increase the modularity. 66

3.8 Example of a node misclassification problem related with recursive algorithms that optimize the modularity. First we analyze the network and we obtain the two partition depicted in the right picture. The circled node is classified in the red partition since it provides higher modularity than if it is classified in the green. We cut the network into this two groups and we perform recursively the analysis of the two partitions, obtaining the final four partitions shown in the left figure. In this case we cannot obtain the maximum modularity, since the circled node should be in the green partition to obtain the maximum, but we cannot explore this configuration due to the limitations of the recursive process. 67

3.9 Analytic reductions for undirected networks. In (a) example of a *hair* reduction, (b) example of a *triangular hair* reduction (see text for details). The widespread case of unweighted networks, all weights equal to 1, implies that in the reduction (a), $w'_{kk} = 2$, and in the reduction (b), $w'_{hh} = 2$ and $w'_{hk} = 2$. 71

3.10 Analytic reductions for directed networks. In (a) example of a *hair* reduction, (b) example of a *triangular hair* reduction (see text for details) 72

3.11 Community structure of the FP6 analysis represented as a network. Nodes correspond to communities and links represent collaboration between members of the two connected communities. Diameter of nodes and width of links symbolize community size and number of crossed collaboration. 74

4.1 Left: Example of the traffic flow that goes through five routers that belong to the Internet2 network gathered from the Abilene project (see text for more details). On the right of each time series we show its average flow and the dispersion. 83

4.2 Value of the exponent α versus the time window length P in which averages are performed, for a fixed $\rho_*^{ef} = 1/3$ and different values of the persistence of packets in the network S. The shadowed area highlights the region of P in which the exponent $\alpha = 1/2$ always appears. 85

4.3 Example of two packet injection rates with the same mean but different variability. This has been done by varying the values of ρ and S proportionally 86

4.4 Left: Plot σ versus $\langle f \rangle$ for different realization of ρ and S maintaining its product constant. Right: Plot of the α exponent for $\rho S = 100$. Other values of ρS have produced equivalent results, shifted to a different region of $\langle f \rangle$. 87

4.5 Scaling exponent α as a function of the time service μ, for three different time window lengths, and for $\rho_*^{ef} = 1/3$. Shadowed area highlights the region where congestion starts at nodes with $\rho_*^{ef} = 1/3$. 88

4.6 Map of the topological structure of the Abilene network. This network acts as the backbone infrastructure of Internet 2, connecting a large number of administrative, educative and private corporations. Map downloaded from http://abilene.internet2.edu/maps-lists/. 90

4.7 Scaling relations between σ and $\langle f \rangle$ for the 112 Abilene backbone router interfaces. Analysis performed during (a) two days, (b) one week, (c) one month and (d) two months, finishing all them in November 15th of 2005. The time window length P is fixed to 5 minutes. 90

5.1 Probability of network splitting in two or more connected components as a function of the fraction of removed nodes p. Inset, relative average path length as a function of p. 96

5.2 Effects of the node removal on the normalized maximum capacity of scale-free and random networks when using a (a) routing protocol based in shortest paths and (b) routing protocol based in random walks. Dotted lines present the analytical approach using equation 5.5 and experimental data from table 5.2. 99

5.3 Effects of the node removal on the normalized maximum capacity of a 2D regular lattice networks when using a routing protocol based in shortest paths and random walks. In the inset we plot the evolution of the relative importance of the node with maximum betweenness as a function of p. 102

5.4 Example of a routing protocol with a certain knowledge radius r. When $r = 0$, node A does not know nothing about the location of other nodes and sends all the packets using a randomly chosen link. When $r = 1$, node A can send directly the packet to nodes in blue (B-F), and otherwise it send the packet randomly. When $r = 2$, know how to reach (B-I) and if $r = 3$ the node know the position of the rest of the nodes, and can send the packets using the shortest path. 103

5.5 Effects of the node removal on the normalized maximum capacity of the ER (right) and SF (left) networks when using a routing protocol that depends on the knowledge radius of the network. The values of r range from 0 (Random Walk protocol) to the Diameter of the network (Shortest path protocol). 104

5.6 Effects of the node removal on some dynamical properties of a communication process as a function of the number of removed nodes p: the average number of steps that one packet performs before exiting the network $\langle D\rangle(p)$ and the maximum algorithmic betweenness normalized $\alpha^*(p)$. The top three figures refer to the Scale-Free network, and the bottom refer to the ER random network. 105

5.7 Comparison between the values of $\rho_c(p)/\rho_c(0)$ obtained using simulations and the values obtained using equation 5.4 and the data of figure 5.6. We observe that the numerical approach provides a good fitting of the behavior of the onset of congestion. 106

5.8 Dynamical robustness versus topological robustness. The dashed line delimits the probability of splitting the network. The crossing between the relative congestion ratio $\rho_c(p)/\rho_c(0)$ and the dashed line determine the change of the dominant effect between both processes. Before this point splitting dominates, beyond this point congestion dominates the communication process after random failures. 107

A.1 Evolution of the number of nodes (a) and of links (b) of the Autonomous Systems topology maps. The nodes and links are shown in a logarithmic scale. 118

A.2 Geographical distribution of the Internet AS nodes in the inferred maps by NLANR in 1997 (c) and by CAIDA in 2007 (d). We see that the growth of the AS network has been centered in the most developed countries. 119

A.3 Evolution of the maximum degree (a) and average degree (b) of the analyzed AS topology maps. Accumulated degree distribution of the two CAIDA snapshots of May 2007 (c) and evolution of the exponent of the accumulated degree distribution (d). Average degree of the nearest neighbors as a function of k of the two CAIDA snapshots of May 2007 (e) and evolution of the scaling exponent between the degree and the average degree of the nearest neighbors (f). 121

A.4 Evolution average shortest path length (left) and of the clustering coefficient (right) of the AS topology maps. 122

A.5 Evolution of the value (left) and size (right) of the maximum k-core. The fluctuations are a consequence of the sampling process used to infer the maps by the mapping projects. 123

C.1 Exemple de l'estructura de comunitats d'una xarxa. Podem classificar els vertexs en grups on el nombre de links cap a membres del mateix grup es més gran que el nombre de links amb la resta de nodes de la xarxa. 135

C.2 Dreta: Estudi de la precisió dels algorismes quan els sometem a xarxes amb una estructura de comunitats prefixada. Com més gran es el valor de z_{out}/z_{tot} més difícil resulta trobar les comunitats. Esquerra: Taula on es resumeix com el cost computacional d'alguns dels algorismes descrits en la tesi escala amb el nombre de nodes del sistema n, el nombre de links m i el grau mig $\langle k \rangle$. La correspondència entre etiquetes i algorismes està explicada al capítol 2 de la tesi. 137

C.3 Exemple del funcionament del nostre algorisme de detecció de comunitats quan analitzem la xarxa Zachary, una de les més utilitzades per a comprovar la precisió de la majoria de mètodes. Gràfic superior: Estat dels nodes de la xarxa despres de la inicialització aleatoria en dos grups i després en cadascun dels moments en que l'algorisme va tallant la xarxa recursivament. Gràfic inferior: Evolució del valor de la modularitat en cadascun dels passos del procés evolutiu. Les barres de separació signifiquen que hem arribat a un estat estacionari i per tant procedim a subdividir el graf en els talls que observem a la part superior. 139

C.4 Dreta: Exemple de dos sistemes que tenen el mateix tràfic mitjà però diferent variabiliat. Esquerra: Transició que es produeix en l'exponent d'escala α a mesura que anem introduint més variabilitat en el sistema. 140

List of Tables

1.1 Assortative coefficient of some complex networks studied by Newman in his article (Newman, 2002). A network is said to show assortative mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections. Social networks usually display assortative mixing ($r > 0$), while technological and biological networks are usually disassortative ($r < 0$). For a detailed explanation of the coefficient and the origin of the datasets see the referred article. 7

1.2 Clustering coefficient C and average path length L of several real networks. To observe the existence of the small-world phenomena, the values have been compared with a randomized version of the network with the same number of nodes and links. It can be observed that the average path length is similar to the randomized, but the clustering coefficient is orders of magnitude higher in the real networks. This data has been reproduced from Albert and Barabási review (Barabási and Albert, 2002), which also includes detailed analysis of other topological properties and describes the origin of the network datasets. 12

1.3 Average degree and exponents of the scale-free degree distribution for several real networks. For the directed networks the table shows both the scaling exponent for the incoming and outgoing degree distribution. It can be observed that almost all the exponents range between 2 and 3. This data has been reproduced from Albert and Barabási review (Barabási and Albert, 2002), which also includes detailed analysis of other topological properties of this networks and the origin of the networks. 15

2.1 Table summarizing how the computational cost of different approaches scales with number of nodes n, number of links m and average degree $\langle k \rangle$. The labels shown here are used in figures 2.9. 49

3.1 Nodes belonging to one of the resultant communitiess obtained when applying the EO method to the FP6 network. Note that all organizations are related, in some sense, with automobiles. See (Lozano et al., 2006) for a complete analysis of this network. 75

3.2 Results for the optimal partition obtained using the upgraded EO algorithm for several real networks before and after applying the size reduction. We present the number of nodes, modularity, number of communities (#c), percentage of size reduction (%r) and speed-up (SU) of the algorithm after reduction. We also compare the modularities and the number of communities obtained before using the two presented improvements of the extremal optimization algorithm (1) and after adding those improvements (2). 77

5.1 Values of the maximum algorithmic betweenness B^*, the importance of this betweenness in the communication process α^*, the average path length $\bar{D}$, and the average generated load by one node $\bar{L}$ for the scale-free and random networks using SP and RW routing protocols. The value of $S\bar{L}\alpha^*/B^* - 1$ determines the change of the congestion point when removing a fraction of nodes p. 100

Chapter 1

COMPLEX NETWORKS: BETWEEN ORDER AND RANDOMNESS

The rise of the Internet is considered one of the keys that have led the technological revolution of the end of the last century, contributing to the globalization phenomena. Collected data confirm this theory; the total number of elements connected to the Internet has been increasing exponentially every year since 1990, with more than one billion people using its services everyday (see figure 1.1). The amount of traffic introduced by these users also grows between 100% and 1000% per year, moving thousands of Petabytes (2^{50} bytes) of information between computers around the world everyday.

Despite the common misconception that behind the Internet there is a highly engineered design, the truth is that the technological part plays a very small role in this exponential growth; its secret relies on a collection of protocols and technical guidelines that describe how to communicate efficiently in a heterogeneous world of electronic devices. Using these guidelines, independent entities (such as governments or Internet Service Providers) have imposed their own regulations to attach new devices to the network, allowing the creation of an extremely heterogeneous structure that has been continuously evolving during the last 20 years.

Regardless of this lack of centralized control and design, the Internet displays two unexpected interesting properties. First, it is one of the most robust networks that actually exist; regardless of the large number of attacks suffered every day by its components (Gordon et al., 2007), very few incidents have been able to produce a global breakdown of communication[1]. Second, the Internet shows an unexpected efficiency at delivering information between users, despite the huge amount of traffic that is distributed by routers worldwide.

[1] The Morris Worm on November 2, 1988 and the attack against the DNS Servers on October 22, 2002 are two of the most relevant attacks against the Internet that have compromised the integrity of the network.

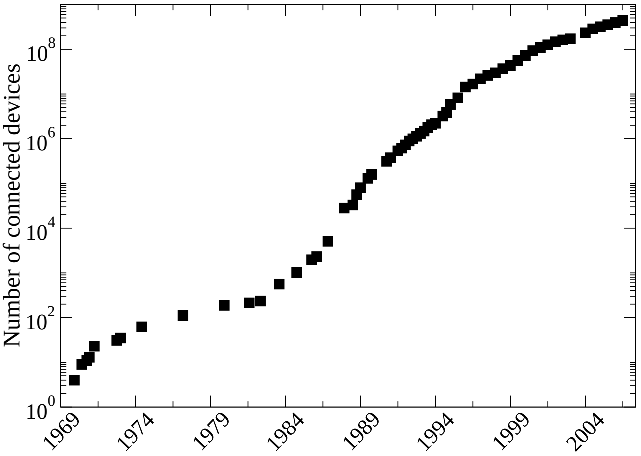


Figure 1.1. Evolution of the number of hosts connected to Internet since the creation of the network in 1969 in a logarithmic scale. Data has been reproduced from the Hobbes' Internet Timeline, which is available for download at http://www.zakon.org/robert/internet/timeline/

From a statistical point of view, the latency or delays suffered by the traveling elements is very low compared to other communication or transport networks (such as highways).

For some of these reasons, Internet is considered as one of the paradigmatic examples of a complex system, which are also present in many natural or artificial environments. Since there is not a standardized definition for complexity, the scientific community usually describes these systems using some common characteristics shared by a large number of them. Like the Internet, complex systems usually display an optimal organization that has emerged without any external control or design. But maybe the most relevant property that all complex systems share is their non-linearity, which results in behavior that cannot be expressed as the sum of the behaviors of their components. This is what sets complex systems apart from complicated systems: in complicated systems, the organization of the different elements is imposed or designed externally. Complicated systems can have a very large number of components, but we are always able to identify the role or function of each element, and using this microscopic information it is relatively easy to infer the macroscopic behavior of the whole system (Amaral and Ottino, 2004). However, the border between complicated and complex is not as clear as it seems, mainly because the lack of a concrete definition for complexity.

Networks are one of the most used representations to describe the underlying connection structure that defines the interactions of the elements of a system. Simple networks are typically represented using graphs, such as lattices or random graphs, which exhibit a high degree of similarity no matter what part is examined, and some dynamical processes based on these structures can be solved analytically. However, most real networks are very heterogeneous and share non-trivial characteristics that cannot be modeled using traditional approaches based on simple networks. These findings led to the development of the “science of networks”, with the main goal of studying complex networks. That is, the group of networks that are “between perfect regularity and total randomness” (Watts and Strogatz, 1998), because in the apparent chaos there is a hidden optimal structure that supports the dynamical processes of a complex system.

The emergence of the new science of networks has been boosted due to three main reasons: the increase in availability of computing resources, the large amount of available networks, and the introduction of tools, measurements, and models that allowed a deep analysis of the networks. Thanks to the computerization of datasets and the emergence of the Internet as a huge repository of information, the scientific community can access to very large amount of network datasets. To understand all these networks, researchers have studied their structural and dynamical properties using three main tools: nonlinear dynamics, statistical physics and network/graph theory. The most interesting point is that the results obtained have been applied successfully in several disciplines, such as computer science, biology, economics or sociology.

In the last five years, some of the main research lines within the science of networks have focused on the study of the interplay between the structure and the collective dynamics of complex systems. Several approaches have proved that there is a bidirectional influence between the topology and the dynamic processes that take place over complex networks. For instance, in the particular case of the Internet, understanding how the topology influences the dynamics of traffic flow provides valuable information on how to design better topologies and more efficient communication protocols.

1. Describing complex networks

Networks are usually described using graphs, composed of nodes (or vertices), that represent the components of the system, and links (or edges) that represent some kind of relationship or interaction between the nodes. The nature of the links can be physical (when there are a real connections between the elements, as in the Internet routers connections) or logical (when they refer to abstract connections, like friendship or collaborations between people).

The study of graphs has its own well established theory in mathematics. Graph theory was set up in 1736 by Euler in his famous problem about the bridges of Königsberg[2]. This theory has solved a large number of problems re-

[2] The problem consists on deciding whether it is possible to walk a route that crosses each of the 7 bridges of the city of Königsberg exactly once.

lated to topological properties of graphs, like the description of their connectivity, Eulerian problems (such as the Eulerian walks), or problems of vertex/edge coloring (Bollobas, 1998).

From a mathematical point of view, a graph G is defined as an ordered pair $G(V, E)$ where V is a set of vertices and E is a set of edges between the vertices $E \subseteq V * V$, $V * V = \{(i,j)|i,j \in V\}$. If the values on the vertices are commutative $(i,j) \in E \Leftrightarrow (j,i) \in E$ the graph is known as undirected. Otherwise we define it as a directed graph (or digraph), and instead of edges we call the links arcs or arrows.

Simultaneously, each edge (or arc) can have an associated a label or value E_w. When the value is 0 or 1 (only informs about the existence or not of the edge) the graph is called unweighted. Otherwise, if $E_w \in \mathbb{R}$ then the links provide extra information about the structure of the graph, and it is called a weighted graph. Other mathematical concepts related with graphs that we will consider in this thesis are the following:

- The order or size of a graph is the number of vertices $|V|$. When the number of edges of the graph is relatively small $|E| \sim O(|V|)$, the graph is called a sparse graph. Otherwise, if it is close to the maximal number of edges $|E| \sim O(|V|^2)$, we classify the graph as dense. When a graph has all the possible edges $|E| = |V|(|V|-1)/2$ it is called a complete graph.

- A subgraph $S(V', E')$ of a graph G is a graph whose set of vertices and set of edges are all subsets of G, $V' \subseteq V$ and $E' \subseteq E$.

- A path is a sequence of distinct vertices V, $\{x_0, x_1, x_2..x_n\}$ where each pair of the sequence is linked ($\forall i, i+1, (x_i, x_{i+1}) \in E$). The length of the path is the number of edges that we have in the sequence.

- A graph is connected if there is a path between any two of its vertices. Otherwise, the graph is disconnected. Each one of the connected parts is known as a component of the graph, and the largest component is usually referred as the Giant Component (GC) of the graph.

The adjacency matrix is the most used representation of a graph. It is a two dimensional matrix with rows and columns labeled as the graph nodes, where each element a_{ij} has a value of 1 (or the weight value) or 0 depending on whether nodes i and j are adjacent or not. If the network is undirected the adjacency matrix is symmetric. And if the network does not have self-loops (i.e. nodes are not connected with themselves), the diagonal of the matrix has only zeros.

However, if one wants to analyze a graph using computational resources, the adjacency matrix is not the optimal representation. On one hand, the amount

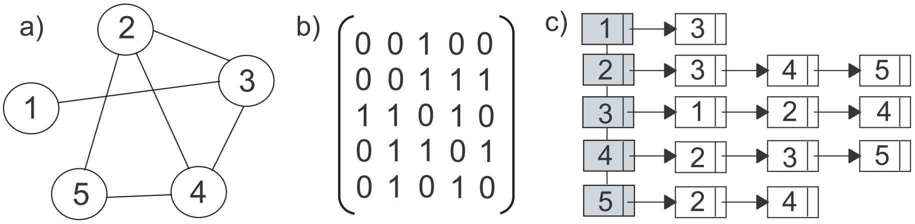


Figure 1.2. a) Example of an unweighted and undirected graph composed by nodes (points) and links (lines). b) Representation of the graph using an adjacency matrix, where the value of $a_{i,j}$ is 1 if there is a link between nodes i and j and 0 otherwise. c) Implementation of the graph using linked lists. The first column are the nodes, and the list that follows each nodes contains the destination of the links that each node has.

of memory needed to store the adjacency matrix delimits the maximum possible size of the graphs (the required space to store a few thousands of nodes overflows the current capacity of computers). On the other hand, since most real networks are sparse, a large portion of the memory is somehow wasted by zeros that are never used in the analysis of the graph.

The development of efficient data structures and algorithms in the 1970's provided better implementations of graphs than storing the whole adjacency matrix, reducing the necessary space and improving the efficiency of the algorithms that deal with the data (Aho et al., 1983). The most used implementation is a data structure known as a linked list, consisting of a vector of nodes that only store the existent relationships between the elements. Using this representation one can store networks of millions of nodes avoiding the capacity problem and without degrading the performance of most algorithms. See figure 1.2 for an example of a linked list.

1.1 Statistical properties of complex networks

The initial goal of the science of networks has been to uncover and characterize network's topology. The first step was the observation of the structural properties of a very heterogeneous group of real networks (biological, social and technological), discovering that most of these networks share similar topological properties, which are not in concordance with the same properties of the traditional regular or random graphs.

To characterize these statistical properties, the researchers developed a set of tools that capture the most relevant topological features. Some of these tools were imported from graph theory and social sciences (such as the degree distribution or the distances between nodes), and they were complemented with the introduction of new specific measurements, like the clustering coefficient introduced in (Watts and Strogatz, 1998) or the assortative mixing introduced in (Newman, 2002).

Using this set of measurements, we can describe the structure of a network at different levels. If we analyze the individual characteristics of each element of the system, we will obtain a description of the microscopic level, and if we consider the properties of the whole system, we will obtain a macroscopic description of the network. For a more detailed description of all these properties see the following reviews (Barabási and Albert, 2002, Dorogovtsev and Mendes, 2002, Newman, 2003b, Boccaletti et al., 2006, da Fontoura Costa et al., 2007).

Degree distributions

The simplest and most studied property of networks is the degree of its nodes k_v, defined as the number of links that vertex v has. If the graph does not admit more than one link between each pair of nodes and the node does not have self-loops, the value corresponds to the number of adjacent neighbors. A first statistical approach is obtained computing the average degree of a network,

$$\langle k \rangle = \frac{1}{|V|} \sum_{v \in V} k_v \tag{1.1}$$

However, if the network is not homogeneous, it is usually more interesting to observe the probability degree distribution, p_k, defined as the fraction of vertices that have a certain degree k. This distribution describes how the degrees are distributed among the nodes of the system, and can be plotted using the following histogram,

$$p_k = \frac{1}{|V|} \sum_{v \in V, deg(v)=k} 1 \tag{1.2}$$

And finally another alternative is to analyze the cumulative degree distribution P_k, which refers to the probability that the degree is greater than or equal to k.

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{1.3}$$

This plot has some advantages over the probability distribution. First, we avoid losing information of data points that fall in the same bin when using a conventional histogram. And second, in the case that the probability distribution has a heavy tail, the cumulative distribution reduces the noise that usually appears in the tail (Newman, 2003b).

Degree correlations

Newman proposed the assortativity as another statistical property of the node degree, measuring the correlation between the degree of adjacent nodes (Newman, 2002). He also pointed out that the models that do not take this property into account will not correctly reproduce many of the behaviors of real systems. The assortative coefficient r is defined as the Pearson correlation coefficient between the degree of connected pairs of nodes. There are three different behaviors that can be determined measuring the value of r in real networks (see table 1.1). When $r \sim 0$, there is no relationship between the degrees of adjacent nodes. This is the typical case of random networks where degree is almost uniformly distributed. Social networks usually have a value of $r > 0$, meaning that highly connected nodes tend to be connected with other high degree nodes. This tendency is referred as assortative mixing, also known as assortativity. On the other hand, many technological and biological networks typically show disassortative mixing (or dissortativity) with $r < 0$, as low degree nodes tend to attach to high degree nodes.

If one wants to get more information about the correlations, another option is to use the relation between the average degree of the nearest neighbors of a node $\langle k_{nn} \rangle$ and its degree k, suggested in (Pastor-Satorras et al., 2001),

$$\langle k_{nn} \rangle = \sum_{k'} k' p(k'|k) \tag{1.4}$$

network	Size	r
physics coauthorship	$52,909$	0.363
biology coauthorship	$1,520,251$	0.127
mathematics coauthorship	$253,339$	0.120
film actor collaborations	$449,913$	0.208
company directors	$7,673$	0.276
Internet	$10,697$	-0.189
World-Wide Web	$269,504$	-0.065
protein interactions	$2,115$	-0.156
neural network	307	-0.163
food web	92	-0.276

Table 1.1. Assortative coefficient of some complex networks studied by Newman in his article (Newman, 2002). A network is said to show assortative mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections. Social networks usually display assortative mixing ($r > 0$), while technological and biological networks are usually disassortative ($r < 0$). For a detailed explanation of the coefficient and the origin of the datasets see the referred article.

where $p(k'|k)$ is the probability that a node with degree k is connected to a node with connectivity k'.

Clustering coefficient

Another important feature that is observed in real complex networks is the existence of a high number of triangles (loops of 3 different edges). This phenomenon is very common in social networks, where reflects the fact that “My friends are also likely to be friends”. To quantify this measure, Watts and Strogatz introduced the clustering coefficient (Watts and Strogatz, 1998) which measures for each node v of the network the proportion of links between its neighbors divided by the number of links that could possibly exist between them:

$$C_v = \frac{2 * (\text{links between neighbors of vertex v})}{k_v(k_v - 1)} \tag{1.5}$$

The clustering coefficient for the whole system is computed as the average value of the clustering coefficient of the nodes.

$$C = \frac{1}{|V|} \sum_{v \in V} C_v \tag{1.6}$$

An alternative definition of the clustering coefficient was introduced in (Newman, 2001b) also taken from social network studies. In this case he proposes to compare the total number of triangles that we can identify in the network versus the total possible number of triangles that can exist. In other words, the value of the clustering coefficient measures the transitivity of the links: if node A is connected to B and C, what is the probability that B and C are also connected?

$$C' = \frac{3 * \text{number of triangles in the network}}{\text{number of possible triangles in the network}} \tag{1.7}$$

Distances and diameters

Although graphs are not usually defined in an Euclidian space, there are some measures of distance that can be defined using the idea of the path introduced previously, counting the number of intermediate steps between two nodes of the network. The most used distance is the average path length L, which measures the average length of the shortest (or geodesic) path between all pairs of nodes of a network:

$$L = \frac{1}{N(N-1)} \sum_{\forall \{i,j\} \in V, i \neq j} d_{ij} \tag{1.8}$$

where d_{ij} is the shortest distance (number of edges) between nodes i and j. Along with L we can also define the diameter of a network as the maximal shortest path length of the whole network, $D = max\{d_{ij}|\{i,j\} \in V\}$. If the network has more than one connected component and there is no possible path between some nodes, we consider both values as ∞.

Centrality measures

Finally, centrality indices are used to measure the relevance of the nodes in the network according to a certain characteristic. The idea also comes from social network analysis, where researchers want to identify the most influential (or central) vertices of a network. For instance, the centrality index of one node can be measured according to its degree (Degree Centrality) or the distance to the other nodes (Closeness Centrality).

Throughout the thesis we are going to work with the betweenness centrality introduced in (Freeman, 1977). This measure plays an important role in the dynamics of communication processes, since it represents how one vertex influences the traffic flow between the other vertices. In other words, it measures the average amount of information that each node has to redistribute. It is defined as the number of shortest paths between all possible pairs of nodes that go through a certain node of the network,

$$B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1.9}$$

where $\sigma_{st}(v)/\sigma_{st}$ is the fraction of shortest paths between nodes s and t that go through node v.

1.2 Network models

The first approaches to model the internal structure of natural systems were mainly based on regular structures like lattices. The work of Erdös and Rényi (Erdös and Rényi, 1959) combined the probability theory with the field of graphs, opening the door for a large amount of alternative research lines and new theories about the structure of real systems. But the recent interest into modeling the structural properties of complex networks arose with the publication by physicists of two canonical works in the field: the small-world model (Watts and Strogatz, 1998) and the scale-free model (Barabási and Albert, 1999). The enormous impact of both publications can be easily understood for two main reasons: first, they provided an empirical demonstration that real complex networks have these non-trivial features that cannot be explained using the previous regular and random approaches. And second, they proved that simple statistical models are able to mimic the structural configuration of real networks with a large degree of accuracy.

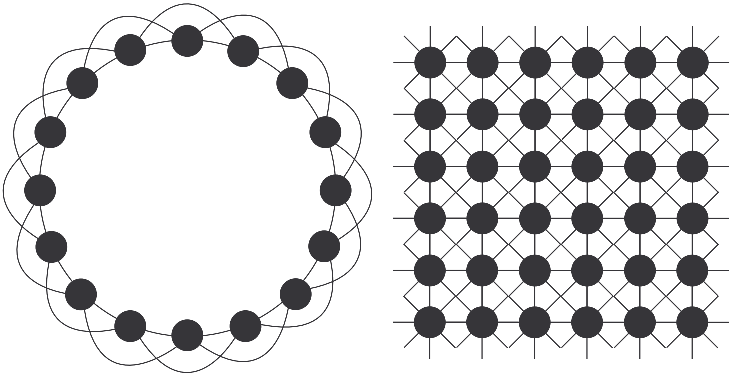


Figure 1.3. Left: Example of a 1D lattice, represented as a ring with periodic boundaries, where each node is connected to neighbors at distance 1 and 2. Right: Example of a 2D lattice in a rectangular grid that also presents periodic boundary conditions.

After publication of these works, network science started to grow in importance and popularity, and a large number of models have been published obtaining different levels of acceptance. These models are mostly used as a platform where mathematical and physical analysis can provide insights about the origin of the structure, and more importantly, to understand the dynamics of the supported complex systems.

Regular networks

Regular networks, and in particular lattices, have been quite popular in physics, since they can describe the organization of the atoms in a crystal or they can be used to discretize some continuum models. Moreover, they are also appealing since in some cases the models based on these structures are exactly solvable. The Ising model or the Potts model are two examples of lattice based models.

The structure of a regular network consists of a set of nodes ordered in a lattice (or other regular structures), all of them connected to all the neighbors that are a fixed distance (see figure 1.3). For simplification purposes, in this dissertation we will only consider regular networks with periodic boundary conditions, where all the nodes share the same exact topological properties. Due to this regularity, one can easily compute the quantities that we have discussed in the previous section, such as the degree distribution, the clustering coefficient or the average path length.

Let us describe these three properties. First, since all nodes have the same degree $k_i = \langle k \rangle, \forall i$, the probability degree distribution follows Kronecker delta function where p_k is 1 for $k = \langle k \rangle$ and 0 otherwise. Second, the average

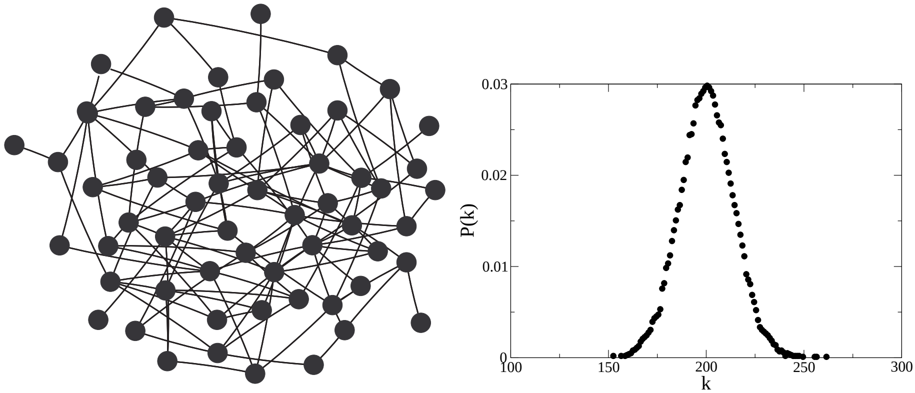


Figure 1.4. Left: Example of a random network generated with the Erdös-Rényi method, with $N = 30$ and $p = 0.1$. Right: Degree distribution of an ER network with $N = 10000$ and $p = 0.02$, with a mean degree $\langle k \rangle \sim 200$.

distance between two nodes in a periodic lattice is usually large, $L_{reg} \propto N^{1/D}$ where D is the number of dimensions of the regular lattice, and increases when we add new elements to the network. And finally, the expected clustering coefficient for this type of networks is almost constant as we increase the size $C_{reg} \sim constant$ due to the periodicity of links.

Random networks

In a classic article of 1959, Paul Erdös and Alfred Rényi proposed a model to create random (or probabilistic) graphs with a fixed number of nodes and links (Erdös and Rényi, 1959). The idea is very simple: select a certain probability p to connect two vertices with one link. Apply the probability to all the possible pairs of nodes of the network and then you obtain an Erdös-Rényi (ER) random network with N nodes and $2p/(N(N-1)$ links (see figure 1.4 left). One of the most interesting aspects of their random model is that as we increase the value of p from 0 to 1, we see how different structural properties emerge. For instance, when the probability p is greater than a threshold $p_c \sim lnN/N$, almost every graph created with the ER method is fully connected.

Let us again describe the same statistical properties that we have analyzed for the regular graphs. The degree distribution of ER networks follows a binomial distribution $p_k = C^k_{N-1} p^k (1-p)^{N-1-k}$. When the network has a large number of nodes, the degree distribution can be approximated using a Poisson distribution with an average degree $\lambda = \langle k \rangle \sim pN$ (see figure 1.4 right).

$$p_k = \frac{e^{-\lambda} \lambda^k}{k!} \tag{1.10}$$

Network	Size	C	C_{rand}	L	L_{rand}
WWW, site level, undir.	$153,127$	0.1078	0.00023	3.1	3.35
Internet, domain level	6209	0.3	0.001	3.76	6.18
Movie actors	$225,226$	0.79	0.00027	3.65	2.99
LANL coauthorship	$52,909$	0.43	1.8×10^{-4}	5.9	4.79
MEDLINE coauthorship	$1,520,251$	0.066	1.1×10^{-5}	4.6	4.91
SPIRES coauthorship	$56,627$	0.726	0.003	4.0	2.12
NCSTRL coauthorship	$11,994$	0.496	3×10^{-4}	9.7	7.34
Math coauthorship	$70,975$	0.59	5.4×10^{-5}	9.5	8.2
Neurosci. coauthorship	$209,293$	0.76	5.5×10^{-5}	6	5.01
E. coli, substrate graph	282	0.32	0.026	2.9	3.04
E. coli, reaction graph	315	0.59	0.09	2.62	1.98
Ythan estuary food web	134	0.22	0.06	2.43	2.26
Silwood park food web	154	0.15	0.03	3.40	3.23
Words, cooccurence	460.902	0.437	0.0001	2.67	3.03
Words, synonyms	$22,311$	0.7	0.0006	4.5	3.84
Power grid	$4,941$	0.08	0.005	18.7	12.4
C. elegans	282	0.28	0.05	2.65	2.25

Table 1.2. Clustering coefficient C and average path length L of several real networks. To observe the existence of the small-world phenomena, the values have been compared with a randomized version of the network with the same number of nodes and links. It can be observed that the average path length is similar to the randomized, but the clustering coefficient is orders of magnitude higher in the real networks. This data has been reproduced from Albert and Barabási review (Barabási and Albert, 2002), which also includes detailed analysis of other topological properties and describes the origin of the network datasets.

The average path length of a random network is $L_{rand} \propto lnN/ln\langle k\rangle$, which for a fixed $\langle k\rangle$ increases very slowly compared to the size of the network (and also much slower than a regular network with the same number of nodes and links). This value is in concordance with the distance observed in real networks (see table 1.2).

When looking at the clustering coefficient, random networks also exhibit a completely different behavior than regular networks. Since random networks have no internal defined structure, they usually have a very low clustering coefficient $C_{rand} \sim \langle k\rangle/N$ (tends to zero as the network is more sparse). This value is also small when compared to the expected ones of real networks (see table 1.2), proving that complex networks are far from randomness.

Small-World : *"Six Degrees Of Separation"*

S. Milgram, a social psychologist, performed an experiment in 1963 to study the distance between two people living in the US (Milgram, 1963). The experiment consisted basically in sending letters between two unknown people, with the instructions of passing the letter to personal acquaintances who they thought might be able to reach the destination target. The average number of intermediate people who received the letter and forwarded it to another was 6, which lead to the idea that between each two people in the US (and afterwards in the entire world) there are the famous “Six Degrees of Separation”. 40 years later the result was reproduced again by Watts and his team. They recreated the same experiment but on a larger scale, using e-mail messages that needed to be forwarded between users around the world. They found again that the average number of intermediaries was around six (Dodds et al., 2003).

From a social network analyst point of view, perhaps the most interesting result of Milgram’s work was that he empirically proved how people were much closely connected than expected. This phenomena was named the small-world effect. Watts and Strogatz redefined the concept of small-world to include those networks that, independently on their size, share two common characteristics: a very short average path length and a high clustering coefficient (Watts and Strogatz, 1998).

The most used technique to check if a network meets the small-world condition is to compare it against a randomized version of the same network maintaining the number of nodes and edges. As we can see in table 1.2, networks that present the small-world characteristic will have a similar average path length than the randomized ones, but its clustering coefficient is much larger. The reason behind this phenomena can be explained in the framework of social networks: usually people share a large number of friends, which gives us the high clustering coefficient. Additionally, each person has a few friends who are far away in distance (e.g. living in other countries), which are represented by connections that reduce drastically the distance between any two people, and therefore, the average shortest path length.

Beyond social networks, this characteristic has also been found in biological (e.g. the neuronal network of the worm *Caenorhabditis elegans*) and artificial networks (e.g. the power grid of the US or the Internet). In these cases, the small-world appears as a consequence of maximizing the functionality of these systems: the clustering provides high redundance (and therefore higher robustness) and the shortcuts improve the efficiency to transmit any signal between two points of the network.

In their 1998 Nature paper, Watts and Strogatz proposed a simple model (from now on WS model) to create a small-world network. Starting from a regular ring lattice, link each node with a fixed number of neighbors. With a certain probability p, rewire some of the links to a random chosen node of the network, without altering the number of vertices or edges in the graph (to represent the long distance friends). For a value of $p = 0$ we recover the regular lattice and, as we increase the value of p, the graph loses its regularity until $p = 1$ where we recover an ER random graph (see figure 1.5 a). Between

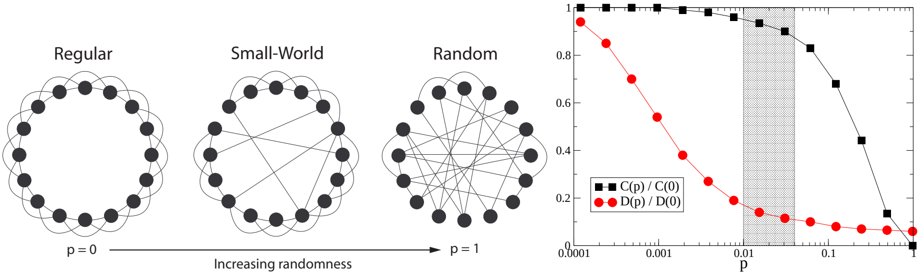


Figure 1.5. Left: Random rewiring procedure introduced by Watts and Strogatz to create a network with high clustering coefficient and low average path length. Right: Normalized average path length L and clustering coefficient C as a function of the rewiring probability p. The shadowed area correspond to a range of values of p that have the small-world property, high clustering and low average path length.

this two values, we can identify a certain a range of p that provides a network that fits into the description of small-world (see figure 1.5 b). Other models have extended the method of Watts and Strogatz, most of them are covered in (Newman, 2000).

Scale-Free : *"The Rich Gets Richer"*

A common feature that the WS small-world model shares with the ER model is that their degree distribution is Poisson, with a well-defined average degree that decays exponentially. After observing a large number or real system, some scientists realized that many real networks does not display this type of connectivity (Faloutsos et al., 1999, Barabási and Albert, 1999). Instead, the degree distribution of real complex networks displays an absence of a characteristic degree, having a few number of highly connected nodes and a large number with a very low degree. These type of networks are known as scale-free networks, and are characterized by a degree distribution with a power-law tail,

$$p_k \propto k^{-\gamma} \tag{1.11}$$

meaning that the probability to choose a node with degree k decays as a power of the degree with a characteristic exponent γ. This exponent is usually in the range $2 < \gamma < 3$ (see table 1.3 for exponents of some measured networks). In other words, the networks with a scale-free degree distribution have a large number of nodes with a very low degree, and a few nodes with large degree, usually having orders of magnitude between the maximum and minimum values.

But the main breakthrough of Barabási and Albert was the introduction of the preferential attachment mechanism, which resembles the work by Herbert Simon in 1955 about "cumulative advantages" (Ijiri and Simon, 1977). They

Network	Size	$\langle k \rangle$	γ_{out}	γ_{in}
WWW	$325,729$	4.51	2.45	2.1
WWW	4×10^7	7	2.38	2.1
WWW	2×10^8	7.5	2.72	2.1
WWW, site	$260,000$			1.94
Internet, domain	$3,015$ - $4,389$	3.42 - 3.76	2.1 - 2.2	2.1 - 2.2
Internet, router	$3,888$	2.57	2.48	2.48
Internet, router	$150,000$	2.66	2.4	2.4
Movie actors	$212,250$	28.78	2.3	2.3
Coauthors, SPIRES	$56,627$	173	1.2	1.2
Coauthors, neuro.	$209,293$	11.54	2.1	2.1
Coauthors, math	$70,975$	3.9	2.5	2.5
Sexual contacts	2810		3.4	3.4
Metabolic, *E. coli*	778	7.4	2.2	2.2
Protein, *S. cerev.*	1870	2.39	2.4	2.4
Ythan estuary	134	8.7	1.05	1.05
Silwood park	154	4.75	1.13	1.13
Citation	$783,339$	8.57		3
Phone-call	53×10^6	3.16	2.1	2.1
Words, cooccurence	$460,902$	70.13	2.7	2.7
Words, synonyms	$22,311$	13.48	2.8	2.8

Table 1.3. Average degree and exponents of the scale-free degree distribution for several real networks. For the directed networks the table shows both the scaling exponent for the incoming and outgoing degree distribution. It can be observed that almost all the exponents range between 2 and 3. This data has been reproduced from Albert and Barabási review (Barabási and Albert, 2002), which also includes detailed analysis of other topological properties of this networks and the origin of the networks.

wisely used the same idea to explain one possible reason behind the scale-free distribution of real networks, establishing some of the bases for the newly-created science of networks.

The preferential attachment mechanism describes how the structure of a complex network evolves using two basic rules: growth and preferential attachment. Starting from a fully connected core of nodes, at each step new nodes are added to the network. Each one of this nodes creates a fixed number of links with the existent nodes following the preferential attachment rule, meaning that the probability to attach to an existent node v is proportional to its degree k_v,

$$p_v = \frac{k_v}{\sum_{i \in V} k_i} \tag{1.12}$$

As a direct consequence of the scientific impact obtained by the publication of the preferential attachment mechanism, a large amount of new models have

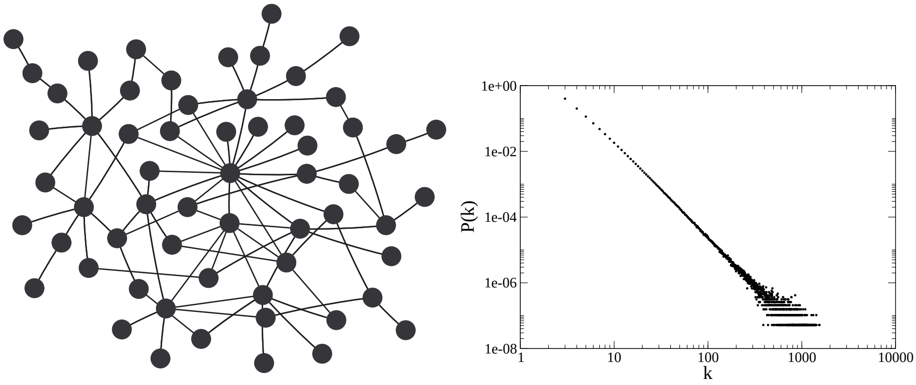


Figure 1.6. Left: Example of a network without a characteristic scale. In this type of networks we usually observe a few high degree nodes that act as hubs, and a large number of peripheral nodes with a low number of connections. Right: Degree distribution of a network generated using the preferential attachment method of Barabási and Albert. The exponent of the power law is ~ 2.8. The average degree of the network is ~ 6 and the maximum degree is ~ 1600.

been proposed to create networks with the same scale-free degree distribution but changing the constraints and rules to add nodes to the network. In (Barabási and Albert, 2002) the authors present an extensive review of these models and the range of γ exponents obtained by each.

Finally, it is interesting to remark that there are multiple ways to obtain the desired degree distribution for a given network without using an evolutionary model. The most used technique is the configuration model (Bender and Canfield, 1978, Molloy and Reed, 1995, Molloy and Reed, 1998). The main idea behind this method is to fix the degree for all the nodes at the beginning, and then try to randomly attach them maintaining the assigned degrees. Using a configuration model one obtains a network where its degree distribution fits accurately the desired one. However, these type of models have been criticized since they are not suitable to represent those systems where growth processes play an important role in the structural evolution of the system (like the Internet or the WWW).

2. Community structure of complex networks

The levels of topological description that have been presented in the previous sections range from the microscopic (degree, clustering coefficient, centrality measures, etc., of individual nodes) to the macroscopic description in terms of statistical properties of the whole network (degree distribution, total clustering coefficient, degree-degree correlations, etc.). Between these two extremes there is a mesoscopic level of analysis of complex networks. In this level we describe an inhomogeneous connecting structure composed by sub-

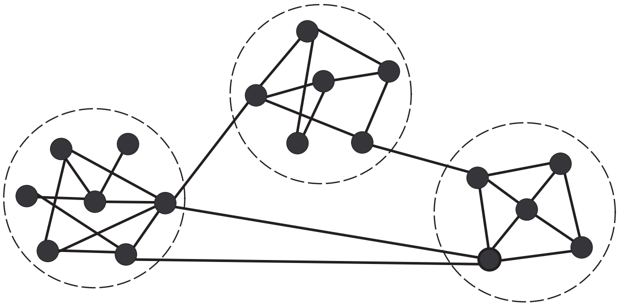


Figure 1.7. Example of the community structure in a complex network. The nodes can be classified into groups where the number of internal links (links between nodes of the same group) is higher than the number of external links (links with the rest of the graph).

sets of nodes which are more densely linked between them than with the rest of the network (see figure 1.7).

The mesoscopic scale of organization is commonly referred as community structure. This concept has been widely used in social sciences (Wasserman and Faust, 1994), where people organize into communities that share common interests, hobbies, or even because they live close to each other. Moreover, this organization can be hierarchical, since for instance a scientist has usually a close relationship with researchers of his group, and at higher levels he has relationships with members of his department or even his university.

The organization of the nodes into communities does not occur only in social networks. Several studies have uncovered the existence of community structures in many different contexts, including metabolic networks (Ravasz et al., 2002, Holme et al., 2002a), banking networks (Boss et al., 2004) or the worldwide flight transportation network(Guimerà et al., 2005). All these studies show that nodes belonging to a tight-knit community are more than likely to have some properties in common. For instance, in the world wide web community analysis has uncovered thematic clusters (Flake et al., 2002, Eckmann and Moses, 2002).

Another group of publications have shown the influence that the community structure has on the dynamical processes that take over the network. Different approaches have been studying the effects on dynamical processes such as synchronization (Arenas et al., 2006b, Arenas et al., 2006a) or emergence of cooperation (Lozano et al., 2007). In all of them the authors show that communities play a key role in the different dynamical processes, explaining phenomena that cannot be understood without the presence of the communities.

The identification and characterization of these clusters of nodes is not a trivial task. A new group of statistical tools have been developed to unravel the existence of community structure in complex networks, providing a large number of approaches that have compared the problem of community identification with some classical problems in physics and computer science. For a more detailed introduction to the problem of detecting the community structure we refer the reader to chapter 3 of this dissertation, since it presents a complete review of the state-of-art of the community detection methods and algorithms, including benchmarks and guidelines to decide which is most appropriate for each problem.

3. Traffic dynamics

Recently, the theory of complex networks has started to cope with the problem of dynamics on networks. After much work devoted to the understanding of the network topology, the physics community has begun to develop models that explain the characteristics of different types of dynamics on complex networks. In this dissertation we will focus on the study of the traffic dynamics of communication processes, trying to understand what are the dynamical properties of the traffic flow between the elements of a complex network. This understanding will help us to design better infrastructures and rules to cope efficiently with the growing demands of traffic volume.

Communication networks provide a background infrastructure that allow a continuous movement of elements, such as information packets in the Internet, electricity in the power grid network, cars in a road, or passengers flying around the world. From a physical point of view, all of these processes can be described using an out-of-equilibrium system of particles that oscillate between different dynamical phases. The first physical approach to study traffic flow dynamics is found in the works of J. Lighthill and G. Whitham who, using fluid mechanics to describe the interactions of the cars in a highway, tried to uncover the reasons behind road congestion (Lighthill and Whitham, 1955). Since then, a large number of different models have been proposed to describe and study traffic flows, which can roughly be divided into macroscopic and microscopic ones depending on the level of description used.

On one hand, microscopic models investigate the behavior of the elements in a concrete part of the network (like cars in an intersection). Each element is considered as an individual entity and usually has an associated differential equation that describes its behavior. The most used microscopic models are based on cellular automata (Maerivoet and de Moor, 2005), which discretize the space of the system into cells and then study how the particles move between the cells. Cellular automata models are numerically very efficient and they have the ability to reproduce a wide range of traffic phenomena, but they lack the accuracy of the time-continuous car-following models.

On the other hand, macroscopic models examine the dependencies between traffic volume, congestion and fluctuations, looking for patterns and other large-scale properties. In this case the system is reduced to a coarse-grained view where the particles are considered equivalent and governed by the same rules (there is not an individual treatment for each element). The difficulty of studying this level is higher than the microscopic level, mainly due to the limitations of storing long time series of data. But thanks to the increasing capacity of the computational resources, we can access to larger time series from multiple elements of the network simultaneously, allowing us to perform a more detailed macroscopic analysis.

3.1 Traffic dynamics on complex networks

In the last decade two factors have increased the attention of scientific community into the study of traffic flows. First, the emergence of complex systems theory has provided the necessary background to characterize the complex behavior of traffic dynamics. And second, the rise of Internet as a worldwide communication network has attracted the attention of researchers that study traffic flow, since they have a huge playground where they can test their theories.

The main goal of the study of traffic on complex networks is to understand the interdependencies between the dynamical parameters and the relevant topological properties. Some stylized models of traffic flow in complex networks (Guimerà et al., 2002b, Tadic et al., 2004, Zhao et al., 2005, Singh and Gupte, 2005, Goh et al., 2005) can be used to gain intuition about dynamics on complex networks, and to determine the leading parameters of the dynamic processes related to the network topology. These models simplify the communication process to the basic elements, using three dynamical parameters to model the traffic flow: the rate of which new packets enter the system, a routing protocol to describe how to distribute the traffic, and a queueing system that represents the limited capacity of the nodes.

The main results obtained up to now concerning traffic flow in complex networks are related to the determination of the bounds for this flow to become congested as a function of the previous parameters. The phenomena of the congestion usually appears in a network when a parameter exceeds a threshold value, provoking a phase transition from the free flow regime to a congested state (Fukuda et al., 1999).

Since the efficient performance of a communication process is a function of the ability of the system to avoid congestion, a large amount of publications have studied the effects of the topology (scale-free degree distribution, small-word, ...) on the onset of congestion (Moreno et al., 2003, Goh et al., 2001). Some studies have proposed optimal network topologies to maximize the amount of information that can be moved over the network without reaching the congestion threshold (Guimerà et al., 2002b, Barthelemy and Flammini, 2006). And another group of publications have proposed new dynamic routing protocols that redirect the traffic when the network reaches congestion (Echenique et al., 2004, Sreenivasan et al., 2007).

A second aspect that has also received a lot of attention is the study of traffic fluctuations. The analysis of the autocorrelations in long time series of data reveals that traffic flows in general, and Internet in particular, display non-stationary behaviors: burstiness across multiple time scales, long range dependencies and self-similarity (Leland et al., 1995). At a macroscopic scale, the fluctuations of the traffic of a system have been characterized as the variability around the mean for all the nodes. Recent studies have confirmed that there is a scaling relationship (de Menezes and Barabási, 2004a), and the exponent of this scaling provides information about the internal and external nature of the fluctuations of the traffic (de Menezes and Barabási, 2004b).

4. The Internet viewed as a complex system

The issue of uncovering and modeling the real structure of the Internet is considered one of the most challenging and attractive open problems of complex networks. Despite all the efforts done by network researchers and computer engineers, we are still unable to see a fully detailed map of its topology. The main reason behind this problem is the lack of a central authority that controls the evolution of Internet. Most part of this infrastructure belongs to private companies that do not share their connection maps, and the only option for researchers is to try to infer them.

Behind Internet there is a multilayered structure, which is a consequence of the different layers of the TCP/IP stack. This adds more complexity into the task of mapping the network, since the meaning of network topology and traffic depends on one's choice of analysis. For instance, we can create network maps of physically connected devices looking at the physical layer, a map of logically connected devices looking at the Internet Protocol layer, or even create a connectivity map of the applications that use this infrastructure, such as the World Wide Web or the Peer-to-Peer (P2P) networks.

Throughout the rest of the thesis we will only focus on the analysis of the structure at the network level. This level can also be subdivided into three different sub-levels of description, obtaining three hierarchical coarse-grained pictures of the network level map:

- User level: here we consider as nodes all the electronic devices connected to the network that can send and receive information. The creation of a global map at this level is a very difficult task due to two reasons: the large number of elements (which is estimated actually around 500 million) and the continuous changes that the topology suffers (e.g. mobile devices change their connection point continuously). Therefore, this level of analysis is only used for statistic purposes like studies of the penetration of the Internet in different geographical areas.

- Internet Router (or IR) level: in this case the nodes represent routers and the links indicate one-hop connectivity between routers (which do not necessarily involve the existence of a physical link between them). In the last measurements, it is estimated that this level is composed by around 200,000 nodes connected by more than 600,000 links.

- Autonomous Systems (or AS) level: an AS is a group of routers and networks managed by a single organization. Usually they are controlled by the Internet Service Providers and public organizations. At this level, the links represent business agreements between two corresponding ASs to exchange traffic between them. The size of the network is estimated around 25,000 nodes and 70,000 links.

4.1 Discovering Internet Topology at the AS level

One of the key properties of the Internet network topology is that it continuously grows in time. New connections are created constantly to maintain the global efficiency of the network as new users join the network. At the AS level, these new connections are mainly guided by economical and technical constraints, since the ASs tend to optimize the economic profit of the infrastructure. On one hand, the growth of the internal structure of each ASs is governed by their own rules. Usually they seem to be engineered to maintain their efficiency at a low cost. This results in small-world like networks, with very short average path lengths and a high local clustering (Govindan and Radoslavov, 2002). On the other hand, new connections between different ASs are established in business relationships, creating what is known as the “Internet Ecosystem” (Norton, 2004). This connections can be peer-to-peer relationships, when they agree to interchange traffic between them, or customer-provider relationships, when one of the ASs provides access to the other one (Gao, 2000).

Due to the strong competition that there is in this market, many ASs do not share the information about their internal structure and their business agreements (e.g. to protect the privacy of their clients or for fear of loosing some advantage), making difficult the task of obtaining a complete detailed map of the complex structure of Internet. However, many important advances have been achieved in the last ten years trying to figure out the AS structure using reverse engineering techniques. The problem is that the techniques are not 100% effective, and usually only provide a partial snapshot of the network.

The initial steps into discovering the topology were performed by computer engineers, who developed a set of analysis and measurement tools to reverse engineer the structure of the network at its different levels. The main source of Internet topology data comes from the Oregon Route-Views Project[3]. This project has been collecting and storing a large number of BGP[4] routing tables since 1997. From this snapshots we can extract all links between the routers and, after some filtering (Chang et al., 2001, Andersen et al., 2002), we can infer topological maps of the AS and IR level.

The topological maps obtained using this methodology have received a large amount of criticism, since BGP data suffers from several limitations. The maps are typically of different quality, sometimes containing errors and ambiguities that depend on the collecting and inferring processes, and the period used to get the data. Moreover, it is known that Internet has some ’dark matter’ which is undetectable using this type of techniques. The number of missing links estimated in the AS maps is between 35% and 50% in the known databases, which mainly are ASs peer-to-peer links (Cohen and Raz, 2006). To solve these problems, new projects propose to discover the Internet topology from a more active point of view. The two most important are the skitter project (Huffaker et al., 1998) developed by CAIDA[5] and the DIMES project[6]. Both projects rely on active sources which ask the network continuously using software probes that are mainly based on traceroutes, a tool that discovers the path between two components of the network. The probes are able to obtain extra routes that are not directly stored in the routing tables, thus obtaining a richer model of the Internet topology than one based on BGP tables. Other techniques, such as WHOIS or looking glasses also provide additional information that can be added to the map, slightly increasing the final number of nodes and links (Mahadevan et al., 2006). The best mapping results up to now have been obtained merging data from different sources, providing progressively new AS maps where we can perform more accurate statistical analysis.

A complementary effort that also receives attention is the visualization of the resulting datasets. The process of drawing the AS resulting maps is a very difficult task, mainly because presenting a network of thousands of nodes in one single snapshot is usually confusing for the viewer. Some methods based on coarse-graining have been used to reduce the visual complexity of the network. Figure 1.8 presents different snapshots of the Internet topology inferred using the techniques described in this section.

[3] http://www.routeviews.org

[4] The BGP (Border Gateway Protocol) is the inter-domain routing protocol that is used in Internet actually. It defines how to distribute the information between the routers belonging to different AS.

[5] http://www.caida.org

[6] http://www.netwdimes.org

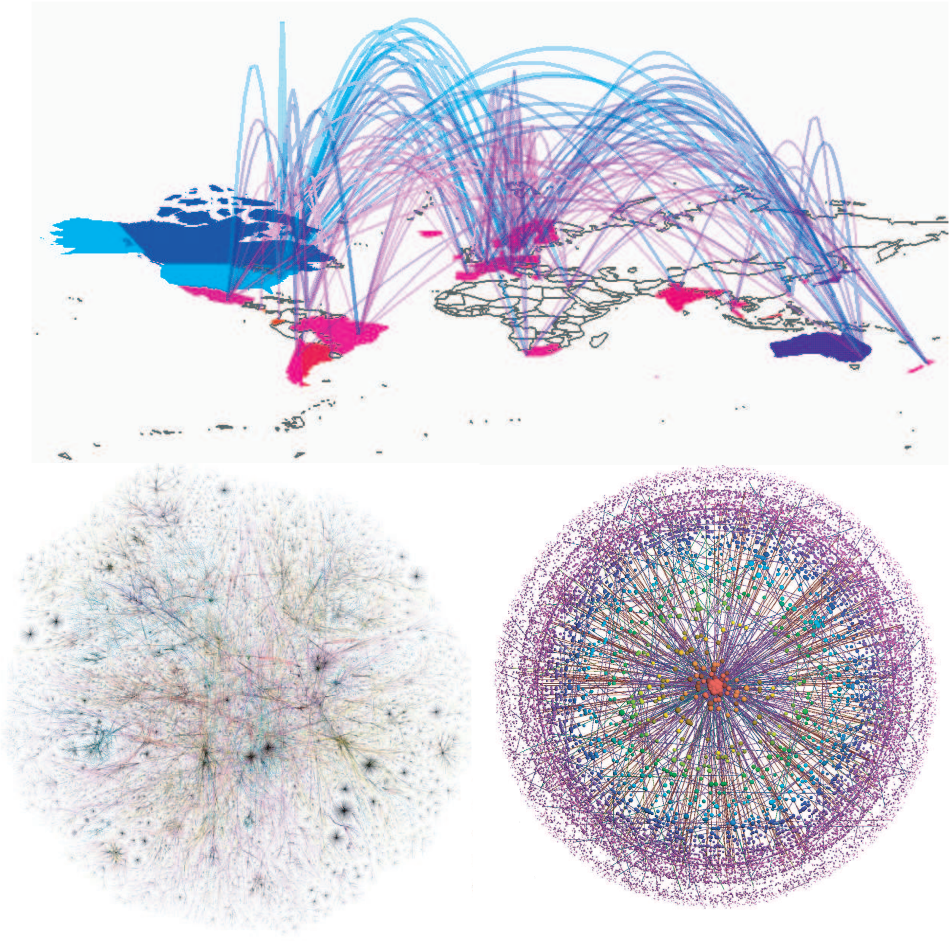


Figure 1.8. The figures present three different views of the Internet topology generated by three different projects. Top: Geographical distribution of the ASs, represented as an arc map, see http://mappa.mundi.net/maps/maps_008/ for more information. The picture has been created by visualization researchers at Bell Laboratories-Lucent Technologies, ©Stephen Eick, Bell Labs. Bottom left: Detailed map of the Internet Router level from the Opte Project http://www.opte.org/. Used under the Creative Commons License. Bottom right: Hierarchical structure of the Internet AS level introduced in (Carmi et al., 2007). The size of the nodes represent their degree and the color their position in the nodes hierarchy.

4.2 Modeling the Internet

Internet models are used to obtain maps which reproduce the structural properties observed in the inferred maps. The most simple models obviate some physical characteristics (such as the bandwidth, router capacity, ...) and represent Internet using undirected graphs. Additional information about the

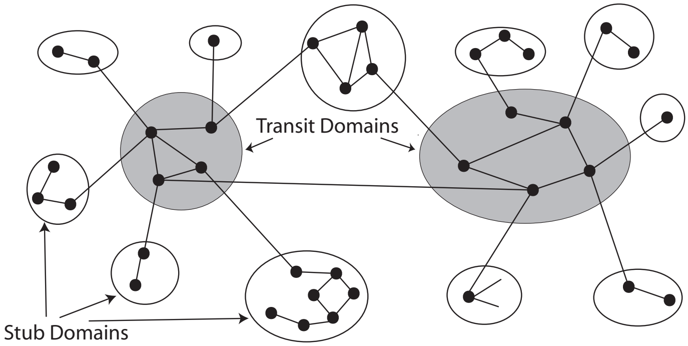


Figure 1.9. Example of an AS Internet topology map generated using the Transit-Stub model. This topology generator creates an Autonomous Systems structure using the two main hierarchical elements of the Internet, the transit and the stub domains.

network can been added to the topological structure by associating information with the nodes and links, and thus obtaining more detailed approximations to the measured networks.

The maps obtained from these models have multiple applications. They can be used as a playground where we can test the efficiency of new routing protocols, to understand certain phenomena like Internet traffic storms (Huberman and Lukose, 1997), or for efficient planning and long-term network design (Yook et al., 2002). Their ability to accurately perform these tasks is directly related to the level of approximation to the real network. Therefore, these models have been continuously changing and evolving to capture the most significant topological properties that are continuously published.

The earliest Internet models were basically stochastic models. The most popular was the Waxman topology generator (Waxman, 1996), which is based on the classical ER graphs with an Euclidean distance constraint to the link probability. Several other models extended this idea, opening a research line focused on representing the local and hierarchical structure of the network. Examples of this type of models are the Transit-Stub model (Zegura et al., 1996) or the Tier model (Doar, 1996), which reproduce the Internet topology as a three-level hierarchy (see figure 1.9 for more details).

However, the canonical work of Faloutsos *et al.* showed that the connectivity of the different nodes on Internet follows a clear power-law distribution (Faloutsos et al., 1999), contrary to the exponential distribution obtained on previous models. The main breakthrough of their work was that they showed the necessity of reproducing the statistical properties to obtain representative maps, opening the door to a new group of ’degree-based’ models. The first of them was proposed by Yook *et al.* based on the preferential attachment mechanism (Yook et al., 2002).

Several modifications on Yook's model have been proposed to capture more and more statistical properties observed in the AS and IR large scale topology, such as the degree correlations, clustering, the maximum degree or the number of loops (Bu and Towsley, 2002, Zhou and Mondragon, 2003). However, all of them have been also criticized since they are merely descriptive and cannot explain the emergence of this properties in the Internet (Willinger et al., 2002). Moreover, it has been proved that topologies with a very different structure can share the same degree distribution and other statistical properties, but when one analyzes the dynamics of Internet routing protocols, their efficiency is completely different (Chang et al., 2006). And finally, a third critic comes from the fact that the Internet is a growing network, and we do not know if the structural properties are enough stable to be reproduced in the models. Therefore, one model that is capable of reproducing a concrete snapshot of the network could not be valid a few months before. In appendix A we give a brief overview of this problem.

The last efforts in this direction are trying to bring together the hierarchical structure while reproducing the statistical properties. An example of this new trend is the "medusa model" (Carmi et al., 2007), where the Internet is described as a nucleus of highly connected nodes surrounded by hierarchical layers of less connected nodes (see figure 1.8).

The future of the Internet modeling still presents interesting challenges, since actual models are imperfect and incomplete. One of the possible ways to improve the models is the introduction into the models the key elements that are behind the growing decisions, such as geographical constraints, user traffic demands or business arrangements. And this must be done without altering the simplicity of the model. A good example of a model based on this ideas is the competition AS model introduced in (Serrano et al., 2005, Serrano et al., 2006). In this case, the growth process of the Internet map is controlled by user and geographical constraints, giving a meaning to the evolutionary growing process. And without imposing any external restriction, the resulting networks reproduce almost every statistical property analyzed in the Internet (including the hierarchical structure), providing one of the best approximations to the measured maps.

4.3 Internet Traffic modeling

The statistical characterization and modeling of Internet traffic has also received a lot of attention from both computer engineers and physicists. The understanding of the physical laws governing the nature of Internet traffic is crucial because of its implications in design, control and speed of the whole network.

The study of Internet traffic is performed both at macroscopical and microscopical levels, measuring parameters such as the amount of traffic that goes through a node, the packet loss rate, or the Round-Trip Time (RTT, the time necessary to travel between two nodes and return). The analysis are performed using long time series of traffic data collected from network hosts and routers.

When modeling network traffic in general, packet and connection arrivals are often assumed to be Poisson processes. However, Leland *et al.* demonstrated that Internet traffic exhibits self-similarity, uncovering the presence of long range dependence in collected traces of packet traffic in local area networks and in wide area networks (Leland et al., 1995). The origins of this self-similarity are still under discussion: on one hand, a group of theories propose that is a consequence of the aggregation of traffic that comes from different protocols, considering that the self similarity emerge as a consequence of the user's actions (Park et al., 1996, Willinger et al., 2002). On the other hand, another group of theories propose a more physical explanation, describing the self-similarity as the consequence of the long-range dependencies that appear when the system is near the phase transition between free and congested regimes (Fukuda et al., 2000, Sole and Valverde, 2001, Valverde and Solé, 2002, Guimerà et al., 2002a).

The characterization of Internet traffic from a large-scale point of view is also a very difficult task, mainly due to the lack of empirical data to prove the different theories combined with the huge complexity of the system dynamics. For the same reasons explained in the previous sections, AS operators do not publish traffic volume statistics or their traffic matrices[7]. The main results up to now are focused on the study of Internet global efficiency, which can be measured using RTT and packet loss rates. A first group of studies have correlated the RTT with the geographical distance and have analyzed the distribution of the RTT, which seems to follow power-law tails (Huffaker et al., 2000, Percacci and Vespignani, 2003). These works have been complemented with the study of the packet loss rate, finding that the probability of having a certain rate of packet loss also follows a power-law distribution (Percacci and Vespignani, 2003).

The results obtained in Internet traffic analysis are in a very preliminary stage. However, all these works devoted to Internet traffic are creating a solid knowledge base about Internet's dynamical behavior, providing the guidelines on how to design the next generation of Internet traffic protocols.

5. Scope of the work

The aim of this thesis is to review and introduce new tools and methods to measure topological and dynamical properties of complex networks. In particular we are interested in two problems, the study of the community structure

[7] Traffic matrices contain the amount of traffic exchanged between two ASs, and additional information such as the delay time or the packet loss ratio

of complex networks, and the analysis of the dynamical properties of a communication process.

Chapters 2 and 3 are focused on the analysis of the community structure of complex networks. In chapter 2 we present an exhaustive review of the community structure identification problem. First we introduce the concept of community structure, the measure of modularity and its limitations. After a complete review of the methods and algorithms available to identify the communities, we present a set of tools to measure the performance of the different methods. Finally, we present a wide range of benchmarks where we compare the efficiency and accuracy of the methods, which can be used to select the best algorithm for a particular problem.

In chapter 3 we introduce the Extremal Optimization method as one of the best alternatives to identify the communities. We explain the physical idea behind the algorithm and how it has been implemented. We also include some improvements that increase the efficiency and the accuracy of modularity based community detection methods, by introducing algorithmic improvements to recursive methods or by reducing the size of the network. Finally we present an exhaustive benchmark of the results obtained by our method when analyzing some of the most used complex networks in the community detection literature.

Chapters 4 and 5 focus on the study of some dynamical properties of communication processes over complex networks. Using a simple traffic model, we have analyzed the changes observed on some properties when we introduce congestion into the network. In chapter 4 we present the scaling of the fluctuations as one statistical measurement that characterizes the behavior of the traffic on a complex network. We analyze how different parameters can explain transitions of the scaling exponent, proving that there is wide range of exponents. We also analyze the particular case of the fluctuations of the Internet.

Chapter 5 introduces the analysis of the dynamical robustness of traffic dynamics, defined as the capability of maintaining the efficiency of the communication when we remove a fraction of nodes of the network. We study how the maximum capacity of one network to deliver traffic changes when we remove a certain fraction of the nodes. We analyze the effect on different network topologies, and using routing protocols that depend on the knowledge radius. We also compare this dynamical robustness with the topological robustness of complex networks.

Finally, the last chapter presents the final conclusions of all the work described in the dissertation and gives some perspectives about how the work can be extended with open questions and new research lines.

Chapter 2

DETECTING COMMUNITY STRUCTURE IN COMPLEX NETWORKS

Numerous studies have tried to explain the relationship between the structure and the functionality of complex systems using the analysis of the structural properties presented in the previous chapter. However, due to the complexity of both the networks and the interactions, this relationship is usually difficult to obtain by looking only at the macroscopic and microscopic levels.

One way to shed light onto this relationship is by studying the intermediate scales of a complex system. It has been suggested that many physical and biological systems display different topological scales (Arenas et al., 2007, Sales-Pardo et al., 2007), and that these intermediate scales affect the behavior of the dynamical processes such as diffusion, communication or synchronization processes (Arenas et al., 2006b, Arenas et al., 2006a, Lozano et al., 2007). Therefore, it seems that the identification and analysis of the intermediate scales of complex networks will enable us to increase our knowledge about complex systems in general.

The main goal of community detection methods is to identify those groups (or communities) of nodes that in real networks share common characteristics or perform similar tasks, but using only information about the topology of the network. The problem of detecting these structures is not trivial and has been the subject of discussion in various disciplines. In real complex networks we typically do not know how many communities there are, but in general there are more than two, making the process more costly than typical bipartitioning problems studied in computer science and statistical physics (Kernighan and Lin, 1970, Fiedler, 1973, Banavar et al., 1987). What is more, communities may also be hierarchical, that is communities may be further divided into subcommunities and so on (Guimerà et al., 2003, Gleiser and Danon, 2003, Arenas et al., 2004).

Despite the difficulties in identifying the optimal division into communities, several methods have been developed and employed with varying levels of success. These methods tackle the problem of community identification from different points of view, by analogy with classical problems such as finding the ground state of a spin glass, the combinatorial optimization of a system or even with a problem of optimal information coding (Mezard et al., 1987, Papadimitriou and Steiglitz, 1997, Shannon and Weaver, 1963). Unfortunately, the problem of having a large number of methods is that the results obtained when analyzing the same network with some of them can provide completely different structures, which raises the questions in the scientific community of which one should they use for a specific problem.

The purpose of this chapter is to present the state of the art on community structure detection methods, providing a set of tools to compare two partitions into communities and also to compare which method performs better. First, we present different definitions for the concept of community, introducing the concept of modularity as one of the keystones of the community detection problem. Next, we summarize the different methods that have been published in the last five years to uncover communities. Then, we introduce a group of benchmarks and measurements that can be used to compare the community structure, and to evaluate the efficiency and the accuracy of the methods. And finally, we give some guidelines that will help to decide which method is the most appropriate for different types of networks.

1. Defining the community structure

Despite the large amount of study in this area, a consensus on what is the definition of community has not been reached. The first approach into the definition of the community structure has its roots in social sciences. This approach is largely (though by no means exclusively) concerned with the effect an individual player has on the network surrounding it and vice versa. As a result, the local properties of networks take a more prominent role in social science research. Some definitions taken from (Wasserman and Faust, 1994) have been used and developed by methods we shall describe later.

Conceptually, the definitions can be separated into two main categories, self-referring and comparative definitions. Central to all such definitions is the concept of subgraph explained in chapter 1. In self referring definitions the basic community definition is *a clique*, defined as a subgroup of a graph containing more than two nodes where all the nodes are connected to each other by means of links in both directions. In other words, this is a fully connected subgraph. This is a particularly strong definition and rarely fulfilled in real sparse networks for larger groups. *n-cliques, n-clans* and *n-clubs* are similar definitions designed to relax the above constraint, while retaining its basic premise. The shortest path between all the nodes in a clique is unity. Allowing this distance to take higher values, one arrives at the definition of *n-cliques*, which are defined as a subgroups of the graph containing more than two nodes where the largest shortest path distance between any two nodes in the group is n. *n-clans* and *n-clubs* are subtle variations of *n-cliques*.

A somewhat different approach to define communities is to compare the number of internal links to the number of external links, coming from the intuitive notion that a community will be denser in terms of links than its surroundings. One such definition, an *LS set* is defined as a set of nodes in which each of its components has more links to other components within the same community. This is the same definition as the *strong definition of community* in (Radicchi et al., 2004). Again the above definition is quite restrictive, and in order to relax the constraints even further, Raddichi *et al.* propose to use the *sum* of links. So a community in the *weak* sense is defined as a set of nodes whose total number of internal links is greater than the total number of links to the outside. This is the most intuitive of all definitions and is the one that is used most, although implicitly.

Self-referring definitions, while useful in characterizing communities which are already known, are not the best choice while trying to find them. The Bron-Kerbosch algorithm (Bron and Kerbosch, 1973) for finding cliques in a network is very costly, running in worst case time that scales exponentially with network size. Comparative definitions, on the other hand, lend themselves much more easily to the search for communities in large complex networks. In a way, comparing the internal structure of a community to the external structure gives rise to a measure of how good a particular partition is, as described in the next section.

2. Detecting community structures

One of the first questions that has been raised in recent years, in the problem of community detection, is how to evaluate a given partition of a network into communities. Using the previous definitions, one can check if different partitions fulfill the strong or weak constraints, but there is no more information to decide if one community structure is better than the others.

A simple approach to quantify a given configuration into communities that has become widely accepted was proposed in (Newman and Girvan, 2004). It is based on the intuitive idea that random networks should not exhibit community structure by definition. Let us imagine that we have an arbitrary network, and an arbitrary partition of that network into N_c communities. It is then possible to define a $N_c \times N_c$ size matrix $\mathbf{e}$ where the elements e_{rs} represent the fraction of total links starting at a node in partition r and ending at a node in partition s. Then, the sum of the any row (or column) of $\mathbf{e}$, $a_r = \sum_s e_{rs}$ corresponds to the fraction of links connected to r.

If the network does not exhibit community structure, or if the partitions are allocated without any regard to the underlying structure, the expected value of the fraction of links within partitions can be estimated. It is simply the probability that a link begins at a node in r, a_r, multiplied by the fraction of links that end at a node in r, a_r. So the expected number of intra-community links is just $a_r a_r$. On the other hand we know that the *real* fraction of links exclusively within a partition is e_{rr}. So, we can compare the two directly and sum over all the partitions in the graph.

$$Q = \sum_r (e_{rr} - a_r^2) \tag{2.1}$$

This is a measure known as *modularity*. Let us consider as an example a network comprised of two disconnected components. If we have two partitions, corresponding exactly to the two components, modularity will have a value of 1. For particularly “bad” partitions, for example, when all the nodes are in a community of their own, the value of modularity can take negative values.

One might be tempted to think that if we search for the maximum modularity in a random network we will found very small values of Q. As Guimerà *et al.* and Reichardt *et al.* show, this in general is not true (Guimerà et al., 2004, Reichardt and Bornholdt, 2006). It is possible to find a partition which not only has a nonzero value of modularity, but that this value can be quite high. For instance, in a random network with 128 nodes and 1024 links we can find a subdivision into communities with a maximum modularity around ~ 0.21. This result raises a new question: how relevant is the partition given by the maximum modularity? Guimerà *et al.* point out that the best way to determine if a modularity is statistically significant is to compare it against a null case, i.e. the randomized version of the same network keeping the degree distribution invariant. The difference between our result and the average value of the null case will help us decide if there is some mechanism behind the network evolution that favors the creation of these clusters (and therefore we can give a meaning to our results) or if the clusters have been created by chance.

There is another issue to take care when considering the partition with highest modularity as the best possible (or the most meaningful) partition into communities. In (Fortunato and Barthélemy, 2007) the authors show a limitation of the modularity to find small communities, instead there is a tendency to combine small communities into larger ones. They show that if a network has L links, it is impossible to identify communities with less than $\sqrt{L/2}$ links by optimizing the modularity, even if these sub-communities are fully connected subgraphs (See figure 2.1). Some new techniques have been proposed recently to override the resolution limit, showing that there is a wide range of community structures at different mesoscales that can lead us to different interpretations of the communities. The first approach has been proposed by Arenas *et*

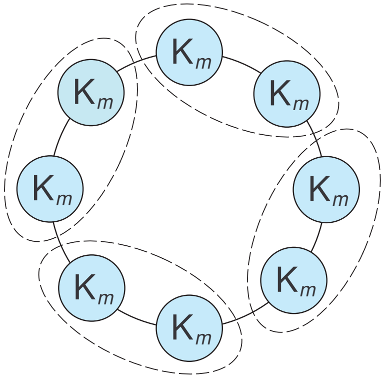


Figure 2.1. Example of the limited resolution of the modularity introduced in (Fortunato and Barthélemy, 2007). The network is composed by identical cliques (complete graphs with m nodes) connected by single links. The methods that optimize the modularity will group the cliques (dotted lines) if there are more than $\sqrt{L}$ cliques, instead of detecting the smaller but highly dense connected groups.

al. and introduces a self-loop of weight r in all the nodes (Arenas et al., 2007). We obtain a graph that maintains the same topological properties than the original (in terms of connectivity), but now we can perform the community analysis at different topological scales adjusting the value of r. This method also allows the identification of the "topological stability" of a given configuration, defined as the range of values of r where we observe this partition; as wider is the range of r where we observe one community, more likely is that this group could have a specific meaning. More recently, Kumpula *et al.*(Kumpula et al., 2007), presented a method to avoid the limitations of the Q-Potts model (see Section 2.7) changing the value of its γ parameter, and obtaining similar results in the number of communities found at different mesoscales of the network.

From here on we will briefly overview the different methods of community identification that have been presented recently, classified into five different sections according to the methodology used to identify the communities. Note that some methods can belong to two or more of this sections, in this case we have chosen the one that we think is closer to the main idea behind the method. First we consider divisive methods that are based on link removal. Then we present agglomerative based methods. In third place we describe methods that try to maximize the modularity. Next we present methods that use the spectral analysis of the network. And finally we present the ones that cannot be classified under the 'other methods' section.

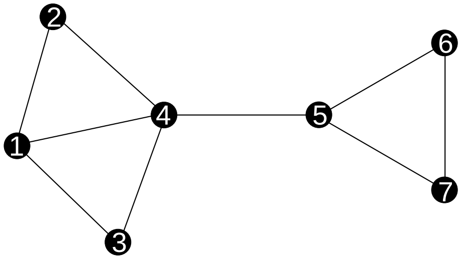


Figure 2.2. Shortest path centrality (betweenness) is the number of shortest paths that go through a link or node. In this simple case, the link with the largest link centrality is that joining nodes 4 and 5.

2.1 Link removal methods

Intuitively, the simplest way to partition a network is to cut some links until the network is no longer connected. Divisive methods do just that. However, cutting links haphazardly is unlikely to give useful results. So, several methods have been proposed to find the most appropriate links to remove, so that the disconnected components correspond to meaningful communities.

Shortest path centrality

One of the first methods to detect communities removes the links depending on their shortest path centrality (Girvan and Newman, 2002). Shortest path centrality measures how central the node or link is in the network, and is computed as the number of shortest paths between pairs of nodes that pass through a certain node or link. Intuitively, links which are most central are also the most “between”, and as such, will act as bridges joining communities together in a connected whole. Removing recursively these bridges should split the network into more densely connected communities, see figure 2.2.

This algorithm is quite sensitive and is one of the few able to detect community structure at all levels. Its major drawback is the computational cost, since calculation of link betweenness requires a computer intensive analysis. It scales with the number of nodes n and number of links m as $O(m^2n)$, which limits the size of the graph one can treat with this method to around 10000 nodes (with current desktop computer technology and some patience).

Current-flow and random walk centrality

In (Newman and Girvan, 2004) the same authors present two other means to detect community structure where the basic method remains the same, with the difference being the way in which the link centrality is calculated. The first approach considers the network to be studied as an electrical circuit, where links are assigned a unit resistance and a particular pair of nodes act as unit

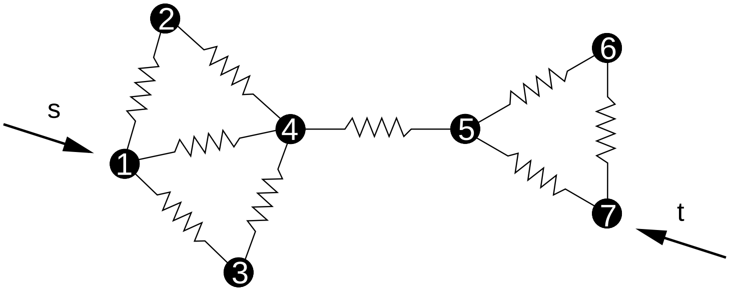


Figure 2.3. Resistor networks and current flow centrality. The links in the network are considered as unit resistances. By choosing a pair of nodes to be a source of unit voltage s and sink t, one can can calculate the current flow through any link using Kirchoff's laws. Summing this value for every pair of nodes gives the total current flow betweenness of a link. In this case the biggest current flow is through link joining nodes 4 and 5.

voltage source and sink. The current flows from source to sink along a number of paths, those with the lowest resistance (shortest path) carry the most current. So the *current-flow* betweenness of an link can be calculated using Kirchoff's laws by summing the value of the current flowing through that link over all pairs of nodes 2.3. In the second approach the network is thought of as a substrate for signals that perform a random walk from a source vertex to a sink vertex. The link betweenness in this case is simply the rate of flow of random walkers through a particular link summed over all pairs of vertices. The authors show that this measure of betweenness is numerically identical to current flow betweenness, but the derivation is different.

Although conceptually interesting, these approaches are computationally costly. As the authors themselves note, and we can see in Sec. 3, the shortest path betweenness outperforms these approaches in both speed and accuracy. Both the resistor network approach and the random walk approach ideas have been developed further by other authors (see posterior sections).

Information centrality

A different divisive algorithm approach was presented in (Fortunato et al., 2004). In this paper they employ the *network efficiency* measure, previously proposed in (Latora and Marchiori, 2004) to quantify how efficient a particular network G is in the context of information exchange. Once a particular link is removed from G, its efficiency is reduced by a measurable amount C^I, or *information centrality*. The idea behind the algorithm is that the links responsible for the largest drop in network efficiency are those that act as bridges between communities. The algorithm is somewhat slower than other divisive algorithms running at ($O(n^4)$), but what it loses in speed it gains in accuracy.

Link clustering

This algorithm, proposed in (Radicchi et al., 2004) is based on the idea that linked nodes belonging to the same community should have a larger number of ’common friends’. In other words links inside communities should be part of a large proportion of possible loops, and links pointing to outside of the community should be included in few or no loops. The algorithm proceeds as in (Girvan and Newman, 2002), but this works removing the links with the lowest ’link-clustering coefficient’ $C^{(g)}$, which represents the fraction of possible loops of order g that pass through a certain link. The algorithm is very fast, since calculating the clustering coefficient can be done with local information only. It is also interesting because it was the first algorithm which contained a definition of community to stop the analysis when a certain condition is fulfilled.

2.2 Agglomerative methods

Instead of starting with the network as a whole and looking for a way to split it into meaningful communities, one can look at the problem from a different perspective. One can start with all the nodes in the network being separate, and use some method to join up, or agglomerate, nodes which are likely to be in the same community.

Hierarchical clustering

Traditional methods for detecting communities in social networks have been based on “hierarchical clustering” (see for example (Scott, 2000) and (Jain and Dubes, 1988)). In general they proceed by calculating a similarity metric for each pair of vertices, representing how close the vertices are according to some property of the network. Such methods have previously been very successful in small scale case studies, particularly when the complexity of the network under study is not great. Recently however, since this method is very fast and scales well with system size, it has been employed to study the temporal evolution of communities in large networks (Hopcroft et al., 2004). Hopcroft *et al.* have studied the CiteSeer citation network (around 250,000 papers) which is intractable with most other methods, demonstrating the ability of hierarchical clustering methods to deal with large data sets.

L-shell method

The algorithm proposed in (Bagrow and Bollt, 2005) consists of creating a shell of nodes of size l. The shell is a subset of nodes, all within a shortest path distance of $d \leq l$ (L-shell) spreading outward from a starting node i. As the shell expands the *total emerging degree*, K_i^l, is measured which is simply the number of links pointing to vertices outside the expanding shell. When the ratio of the emerging degree at step l to that at step $l - i$, $\frac{K_i^l}{K_i^{l-1}}$, is lower than a cut-off value, the algorithm is stopped, grouping all the nodes within a distance l of the starting vertex within one community, and all other nodes are said to be outside.

This algorithm is specially useful when one is concerned with a single community and not the entire community structure, and for this purpose the algorithm is computationally inexpensive scaling linearly with the size of the community under scrutiny.

K-clique method

Another approach proposed by Palla *et al.* (Palla et al., 2005) introduces the idea that communities can overlap. In their definition of community, one node can belong to various “tematic” communities (i.e. one can belong to a scientific group, a family, a sports team, ...), which usually share a certain amount of nodes (see figure 2.4 a). The idea behind this overlapped communities is based on the concept of k-clique communities. A k-clique is a group of k nodes that is a complete subgraph, and a k-clique community is the union of all k-clique that are adjacent (two k-cliques are adjacent if they share $k - 1$ nodes). Searching all the possible k-cliques of the network will provide a result similar to figure 2.4 b.

In terms of accuracy, this method is not comparable with the others presented, since it uses a different definition of community structure. However, it has other interesting applications, i.e. it can be used to observe the relationship

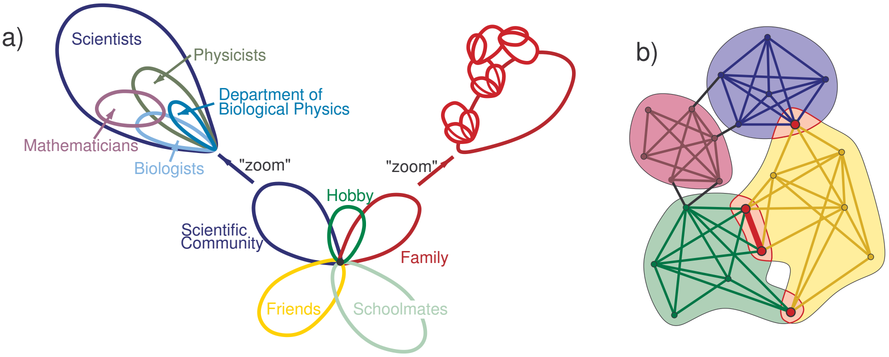


Figure 2.4. a) Overlapping communities around a given node. We can observe that one node can belong to more than community at the same time. The communities can overlap and share more than one node between them b) An example of overlapping k-clique communities at with $k = 4$. The red nodes belong to more than one community. For more information about the method see (Palla et al., 2005)

between the different communities or to determine the communities where a certain node belongs.

2.3 Methods based on maximizing modularity

As described previously, the modularity measure is one way to evaluate quantitatively a network partition. So, as many authors have asked themselves, why not optimize this value directly? The main problem is that the partition space of any graph (even relatively small ones) is huge, and one needs a guide to navigate this space and find maximum values. Here we outline the approaches that have tackled this problem.

Greedy algorithm

In the first attempt at optimizing Q directly Newman takes a greedy optimization (hill climbing) approach (Newman, 2004b). At the start of the algorithm, each node is placed into its own partition. One can then calculate the change in Q should any two partitions be joined. The algorithm proceeds by choosing the pair of partitions producing the largest change, and joining them. This process is repeated until a maximum value of Q is obtained. The algorithm is one of the fastest available, especially when applied using the data structure for sparse networks described in (Clauset et al., 2004). However, while also pretty good at identifying community structure, more recent approaches have achieved even more accuracy (see Sec. 3).

The main drawback of this method is that it tends to favor the creation of large communities at the expense of smaller ones. With a simple modification of the algorithm, Danon *et al.* presented a method capable of identifying heterogeneous communities ensuring that communities of differing sizes are treated equally (Danon et al., 2006), improving the efficiency Newman's method without increasing the temporal cost. Another interesting upgrade of this method has been proposed in (Pujol et al., 2006). Instead of placing the nodes individually at the beginning, they perform a random walk process to reduce the dimensionality of the network. In this initial process, they group the nodes according the number of times that a certain walker have visited them. After this they use the greedy algorithm to optimize the modularity. This modification increases the accuracy while reduces the temporal cost compared with Newman's original version.

Simulated annealing methods

Another approach to optimize the modularity measure is to employ simulated annealing methods. It was first proposed by Guimerà *et al.* to study modularity in random networks (Guimerà et al., 2004). The process begins with any initial partition of the nodes into communities. At each step, a node is chosen at random and moved to a different community, also chosen at random. If the change improves the modularity it is always accepted, otherwise it is accepted with a given probability. The process is repeated until we cannot improve the modularity anymore. The algorithm is slower than some of the other methods, but as we present in sec. 3 is the most accurate option up to date.

In (Massen and Doye, 2005) the authors present two modifications of the Monte Carlo sampling method with simulated annealing. Firstly, the algorithm is stopped periodically, or quenched. Then they analyze all the possible node movements and accept the move corresponding to the largest increase of the modularity. The second way to improve the efficiency is using a Basin-Hopping approach, where in each step a series of nodes are moved from one community to another, not just one. In this case, the acceptance criterion is calculated directly from the partition that results at the end of the move. The authors report that this method is slower to run, but is able to find high values of modularity quickly. In case of large networks it requires less computer memory than the other presented, since it doesn't need extra data structures.

2.4 Spectral analysis methods

An alternative representation of a graph other than the adjacency matrix is the Laplacian matrix. If a link exists between nodes i and j, the element $L_{ij} = -1$. The diagonal of the matrix L_{ii} contains the degree of node i, so that the sum of each row and column is equal to zero. Methods which take advantage of algebraic properties of these matrices have been proposed over several decades in many physical and mathematical problems.

Multi dimensional spectral analysis

Taking advantage of the properties of the Laplacian matrix, Donetti and Muñoz present a very nice approach in (Donetti and Muñoz, 2004). The first few non-trivial eigenvectors can be extracted sequentially at minimum cost using the Lanczos method, which can be applied to sparse matrices at minimum computational cost (Golub and van Loan, 1996). The individual eigenvector components, which represent nodes in the graph, can be thought of as coordinates in M-dimensional space, where M is the number of non-trivial eigenvectors considered. The idea is that if two nodes belong to the same community, they are close in this M-space. Once separated in this space, the nodes can be clustered using hierarchical agglomerative methods (i.e. "single linkage" or "multiple linkage"), using both simple Euclidean distance and angular distance. The clustering is stopped at the highest value of modularity obtained, thus detecting the optimal configuration.

This algorithm is reasonably fast but needs *a priori* information on how many vectors need to be extracted to separate the communities properly. In

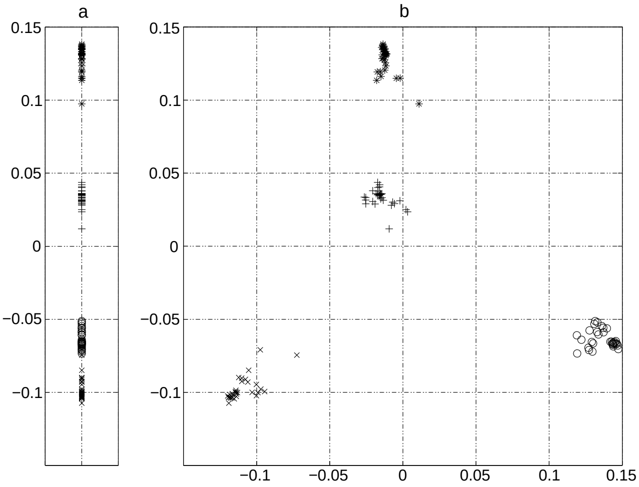


Figure 2.5. a) Components of the first non-trivial eigenvector for a *ad hoc* network with 4 communities (see Sec. 3). b) All communities can be clearly identified when the components of more than one eigenvector are used as coordinates in M-dimensional space where M is the number of eigenvectors used. Here $M = 2$.

terms of sensitivity, the algorithm performs well (see Sec. 3). In the comparison section, we use the aliases DMCS and DMCA for Single Angular and Complete Angular analysis respectively.

Constrained optimization

This method, described in (Capocci et al., 2004) is based on the spectral properties of the simple adjacency matrix as opposed to the Laplacian. The authors recast the costly problem of extracting eigenvectors of an $N \times N$ matrix into a constrained optimization problem. In this way they are able to extract the eigenvectors much faster. As in the previous method this gives information about the location of the different nodes ordered in different groups in an M-space (where M is once again, the number of eigenvectors extracted). To detect the groups that appear, they use a correlation of the average values of the eigenvectors to measure how close two nodes are in this space. Instead of providing a clear cut community structure, this method gives us an idea of how close any pair of nodes is in the context of communities.

Spectral optimization of the modularity

A different approach to detect communities using the matrix spectra has recently been introduced in (Newman, 2006b). The idea is to rewrite the modularity function in matrix terms, and then detect the communities using spectral partitioning methods. The modularity matrix is defined as $B_{ij} = A_{ij} - P_{ij}$,

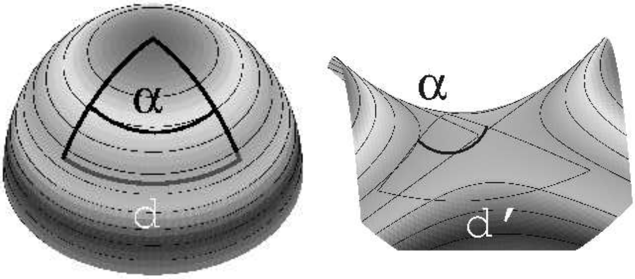


Figure 2.6. How clustering is related to curvature according to (Eckmann and Moses, 2002). For a node i, the shortest path distance between any of its neighbors will be either 1, if the neighbors are linked, or 2, if they are not. The average distance between the neighbors can give a measure of curvature. Positive curvature is depicted in (a) and negative curvature is depicted in (b). Both triangles have sides of length unity, and the angle between the two is the same, but the distances are different.

being A the adjacency matrix and P_{ij} the probability that, maintaining their degrees, nodes i and j are connected in a randomized version of the network.The idea is the following. First we need to compute the leading eigenvector of the modularity matrix. And then, depending on the sign of the values of this eigenvector, the nodes are classified on different communities. The division into communities is performed by recursive divisions into two communities while we optimize the modularity. The total cost of the algorithm scales with O($n^2 \log n$), and the modularity values achieved in some test networks are among the highest.

2.5 Other methods

This section is dedicated to those methods that do not belong clearly to any of the previous classes.

Clustering and curvature

This is one of the first attempts at detecting thematic and functional communities based on clustering (Eckmann and Moses, 2002). The authors use the concept of *curvature* of a node and relate it to clustering. Consider a node i; its neighbors will be separated by a geodesic distance of at most 2. If links exist between neighbors of node i, this distance is unity. The average distance between neighbors of any node, therefore, lies between 1 and 2. This value is directly related to clustering (see (Eckmann and Moses, 2002)). If one assumes that the distance from node i to any of its neighbors is unity, and take the distance between any of the neighbors to be the average, one can indeed think of the node to be in “curved” space, with the amount of curvature depending on the average distance between the nodes, see figure 2.6. The method is based on the intuition that high curvature region of a network will belong to the same community. The authors show that finding connected components of high curvature give a good idea of community structure.

Random walk based methods

In a set of papers, Zhou and collaborators develop a methodology for community detection based on random walks (Zhou, 2003a, Zhou, 2003b, Zhou and Lipowsky, 2004). The authors show that instead of actually performing the random walk on the network, it is possible to calculate the average distance between two nodes algebraically using with the adjacency matrix. From the information contained in the average distances, the authors define which nodes act as global and local attractors[1], and then agglomerate the nodes according a set of rules based on the hierarchy of the attractors.

The authors have proposed some interesting modifications to their original method. First, in a more refined effort, the authors use the average distance measure to define a *dissimilarity index* of any two nodes[2] (Zhou, 2003a). Using the dissimilarity index, the author describes an elaborate method of hierarchical agglomeration of nodes into communities. And more recently they have presented another method based on *biased random walks* (Zhou and Lipowsky, 2004, Zhou and Lipowsky, 2005). Instead of having the walkers performing purely random walks, the walker has a higher probability to jump from a node i to a node which shares the highest number of neighbors with i (essentially biasing the random walker to go down the link with the highest link clustering).

In a similar approach Latapy and Pons (Latapy and Pons, 2004) also employ the intuitive idea that a random walker will get trapped for a longer time in a a densely connected community. They calculate a distance measure between two nodes, and apply an agglomerative method (Ward, 1063), starting with all nodes in their own community, and joining them two by two. The main difference between this approach and the above is that at each step, the distances are recalculated. The two methods have very similar sensitivities, suggesting that recalculating the distances in each step is not crucial, see Sec. 3.

Approximate resistance networks

In a development of the resistor network approach in (Newman and Girvan, 2004) Wu *et al.* present an approximate method, in order to reduce the computational time needed (Wu and Huberman, 2004). The authors select two nodes as source and sink, assign them a fixed voltage, and then approximate the volt-

[1] The local attractor of node i is the closest node (smallest average distance) of its nearest neighbors, and the global attractor, the node closest to all other nodes in the network

[2] For nodes i and j the dissimilarity index is simply the square of the difference between the distance from another node k to i and the distance from k to j summed over all nodes k.

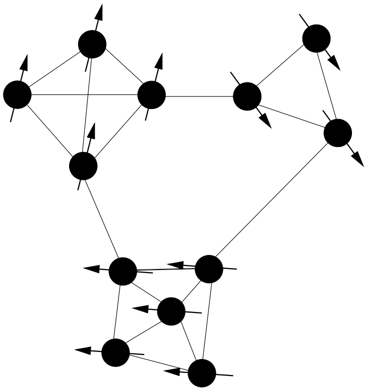


Figure 2.7. The q-Potts model as applied to a small network with communities. Each node is assigned one of q spins. As the Hamiltonian of the system is minimized, the spins in a tightly connected community take equal values, which are different to those of spins located in other communities.

age of the rest of the nodes. The process is performed iteratively, avoiding the costly matrix inversion used in (Newman and Girvan, 2004). Identifying the gaps in voltage values they can split the graph at a particular voltage gap, separating a number of nodes (within a tolerance limit), which must be previously known, from the rest of the network. This process is repeated, randomly choosing pairs of nodes to be voltage sources and sinks. Finally nodes are then bundled together into a community of the expected size using a simple majority rule over the realizations of the algorithm.

This method when employed to identify all communities in a graph is dependent on having a good idea of the sizes of communities one is looking for. In networks of larger size and complexity, this is generally not known, and the algorithm becomes more difficult to apply. However, the method can be employed to identify in linear time the community that any one nodes belongs to, similar to the approach of the L-shell method.

Q-potts model

Another interesting approach (Reichardt and Bornholdt, 2004) detects communities by mapping it to a spin system (Blatt et al., 1996). Here, each node is assigned a spin state between 1 and q, at random. The energy of the spin system is determined using a q-Potts Hamiltonian[3]. The idea is that in the ground state of the system, communities are identified as groups with equal spin values, see figure 2.7. One useful characteristic of this is that it permits

[3] The q-Potts model is essentially an Ising model with q states instead of just two

the detection of communities which are “fuzzy”, or clearly separate from the rest of the network. The method should be fast since one only needs only local information to calculate the Hamiltonian and update the spins. The sensitivity of the algorithm is also good, as we can see in the next section.

Information theoretic approach

One of the most recent approaches to the community detection problem has been proposed in (Rosvall and Bergstrom, 2007). It is based on an information-theoretic framework, where the community detection problem is now treated as an information compression problem. The idea is to reduce the link connectivity of the network (the adjacency matrix) into a more simple description (a module assignment vector and a module matrix). To discover the configuration that provides the best “compression” of the network structure, they maximize the mutual information between the encoded and the global descriptions.

The results presented in their paper show that this method performs better than the others when detecting asymmetric communities. Another advantage is that changing the encoding function we can detect other types of clustering beyond the classical community structure. Similar to the mixture models presented in (Newman and Leicht, 2007), the method is also able to identify partitions where the nodes have similar patterns of connection to other nodes.

3. Comparative evaluation

Thus far we have described several methods to identify the optimal community structure from a wide range of points of view. In this section we would like to present a qualitative comparison for all the methods, but this is not possible as they are very varied, both conceptually and in their applications. Therefore, our main goal is to compare the efficiency and accuracy of as many as possible methods, which will help us write some guidelines on what methods are recommended to analyze different types of networks.

3.1 Accuracy of the methods

One way that has been employed to test sensitivity in many cases is to see how well a particular method performs when applied to *ad hoc* networks with a well known, fixed community structure (Newman and Girvan, 2004). Such networks are typically generated with $n = 128$ nodes, split into four communities containing 32 nodes each. Pairs of nodes belonging to the same community are linked with probability p_{in} whereas pairs belonging to different communities are joined with probability p_{out}. The value of p_{out} is taken so that the average number of links a node has to members of any other community, z_{out}, can be controlled. While p_{out} (and therefore z_{out}) is varied freely, the value of p_{in} is chosen to keep the total average node degree, k constant, and set to

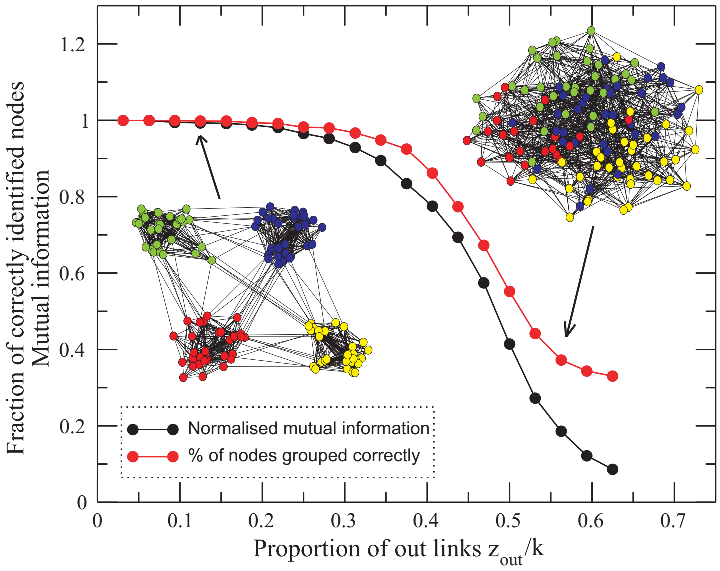


Figure 2.8. Algorithm sensitivity as applied to ad hoc networks with $n = 128$, the network divided into four communities of 32 nodes each and total average degree z_{out} fixed to 16. For low z_{out}/k the communities are easily distinguished. For higher z_{out}/k this becomes more complicated. Both measures of comparing original communities to ones found by the detection method are shown. The normalized mutual information measure is more discriminatory and appears more sensitive to errors in the community identification procedure. The results are shown for Newman's fast algorithm (Newman, 2004b).

16. As z_{out} is increased from zero, the communities become more and more diffuse and harder to identify, (figure 2.8). Since the "real" community structure is well known in this case, it is possible to measure the number of nodes correctly classified by the method of community identification.

In (Newman, 2004b), the author describes a method to calculate this value. The largest group found within each of the four "real" communities is considered correctly classified. If more than one original community is clustered together by the algorithm, all nodes in that cluster are considered incorrectly classified. For example, for the case when z_{out}/k is small, if a method finds three communities, two of which correspond exactly to two original communities, and a third, which corresponds to the other two clustered together, this measure would consider half the nodes correctly classified. As the author notes, this measure is quite harsh, and some nodes which one may consider to be correctly clustered are not counted. On the other end of the spectrum, as z_{out}/k becomes large, and the networks become essentially random networks, this method rewards the identification of smaller clusters found within each of the original communities, which could be misleading.

We suggest that a more discriminatory measure is more appropriate, and propose the use of the *normalized mutual information* measure, as described in (Kuncheva and Hadjitodorov, 2004, Fred and Jain, 2003). It is based on defining a *confusion matrix* $\mathbf{N}$, where the rows correspond to the "real" communities, and the columns correspond to the "found" communities. The element of $\mathbf{N}$, N_{ij} is the number of nodes in the real community i that appear in the found community j. A measure of similarity between the partitions, based on information theory, is then:

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{c_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{2.2}$$

where the number of real communities is denoted c_A and the number of found communities is denoted c_B, the sum over row i of matrix N_{ij} is denoted $N_{i.}$ and the sum over column j is denoted $N_{.j}$

If the found partitions are identical to the real communities, then $I(A,B)$ takes its maximum value of 1. If the partition found by the algorithm is totally independent of the real partition, for example when the entire network is found to be one community, $I(A,B) = 0$.

Both measures of accuracy give a good idea of how a method performs. However, the measure we propose for use here is more representative of sensitivity if the performance is dubious, since it measures the amount of information correctly extracted by the algorithm explicitly. As an example, for small z_{out}, where two original communities are clustered together by the algorithm, this measure does not punish the algorithm as severely, taking into account the ability to extract at least some information about the community structure. On the other hand, for large z_{out}, this method is able to detect that the clusters found by the algorithm have little to do with the original communities, and $I(A,B) \to 0$.

In figure 2.9 we show the sensitivity of all methods we have been able to gather. The percentage of correctly identified nodes is calculated using the method described in (Newman, 2004b), since this is the method employed by the various authors. We can see that accuracy varies in a similar way across the different methods as z_{out} increases and the communities become more diffuse. So, it remains difficult to compare the performance by looking at the methods separately, even with a reference performance.

To summarize the large amount of information, in figure 2.10 we plot the fraction of correctly identified nodes for only three values of z_{out} (6, 7 and 8), corresponding to z_{out}/k = 0.375, 0.4375 and 0.5 respectively, for each method. From this we can see that most of the methods perform very well for $z_{out} = 6$ ($z_{out}/k = 0.375$), and even for $z_{out} = 7$ ($z_{out}/k = 0.4375$) most can

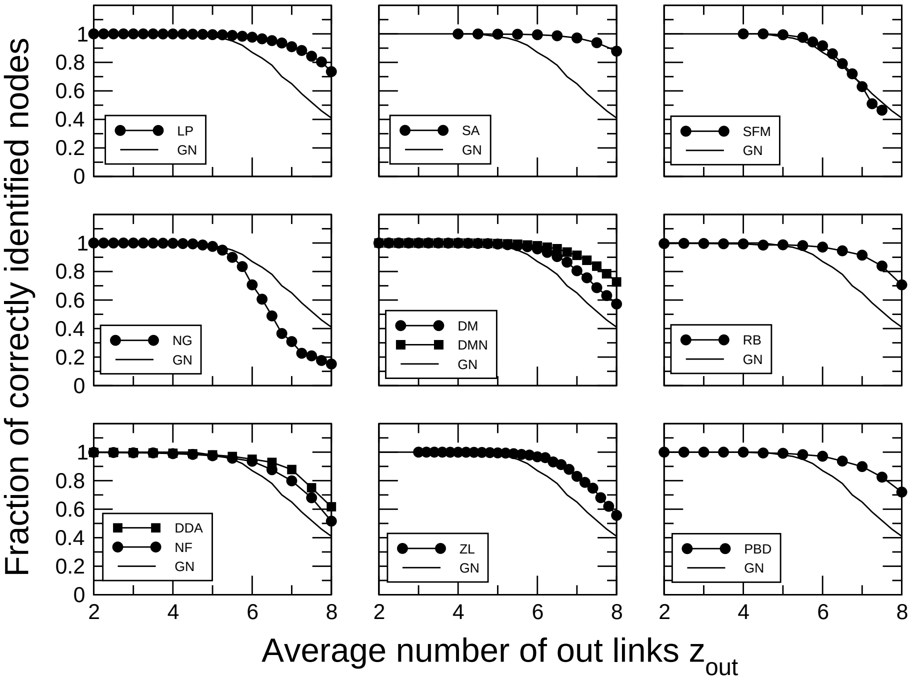


Figure 2.9. Comparing algorithm sensitivity using ad hoc networks with predetermined community structure. The x-axis is the proportion of connections to outside communities z_{out}/k and the y-axis is the fraction of nodes correctly identified by the method measure as described in (Newman, 2004b). The labels here correspond to the different methods and are listed in table 2.1.

identify more than half the nodes correctly. For $z_{out} = 8$ ($z_{out}/k = 0.5$) the SA method is still able to identify more than 80 $\%$ of the nodes correctly.

Although these are the most used reference networks to compare the accuracy of the methods up to date, they have been criticized because they do not reproduce the community structure observed in real networks. In order to expand the scope of these comparisons, some new benchmarks have been proposed that explore the accuracy of the methods when confronted against other artificially controlled networks. Here we present two methods based on small modifications of Newman's networks. The first modification, introduced in (Danon et al., 2006), describes how to create networks to test the effect of size heterogeneity, reproducing the fact that in real networks usually the distribution of community sizes is highly skewed. To generate such networks we need to chose the size of all the communities and a factor that helps us to control their internal and external cohesion (see figure 2.11 left). A second modification provides computer generated networks with a well-defined hierarchical substructure (Arenas et al., 2006b). Now we define the sizes of the hierarchi-

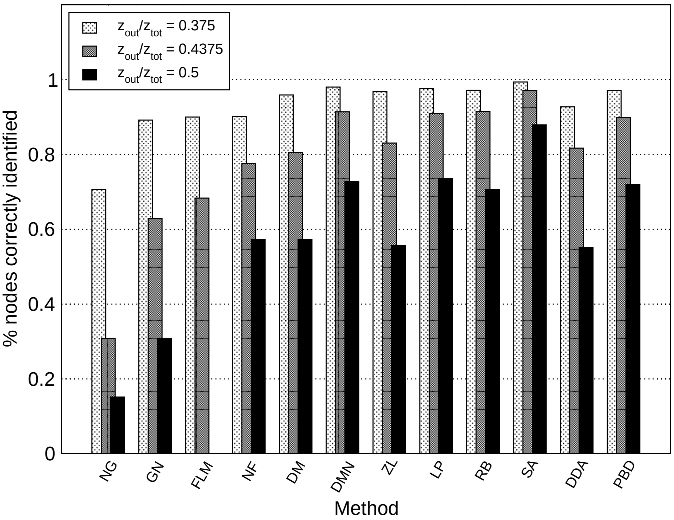


Figure 2.10. The fraction of correctly identified nodes at three specific values of z_{out}, 6, 7 and 8 for all available methods and for networks with fixed $k = 16$. Note that for the FLM method, the data for $z_{out} = 8$ were not available. Here we can see that most of the methods are very good at finding the "correct" community structure for values of z_{out} up to 6. At $z_{out} = 7$ some methods begin to falter but most still identify more than half of the nodes correctly. At $z_{out} = 8$, when on average half the links are external, the SA method is still able to identify over 80 % of the nodes correctly.

cal subgroups and the probabilities to connect two nodes depending on their relation in the hierarchical structure (see figure 2.11 right). These networks provide a good benchmark to test if a method is able to unravel the different mesoscales of the community structure.

3.2 Efficiency of the methods

While accuracy is an essential consideration when choosing a method, it is just as important to consider the computational effort needed to perform the analysis. For some of the approaches described in the literature, we have collected estimates of how the cost scales with the size and/or density of the network. For networks with n nodes and m links, the methods scale between $O(m + n)$ for the fastest, and $O(\exp(n))$ for the slowest as it is shown in table 2.1. Such diversity is due to the heterogeneous approaches taken by the authors. The faster methods tend to be approximate and less accurate, while the slower methods have other advantages. Differences in speed only become important when dealing with larger networks, and for smaller networks we can choose between the more accurate ones.

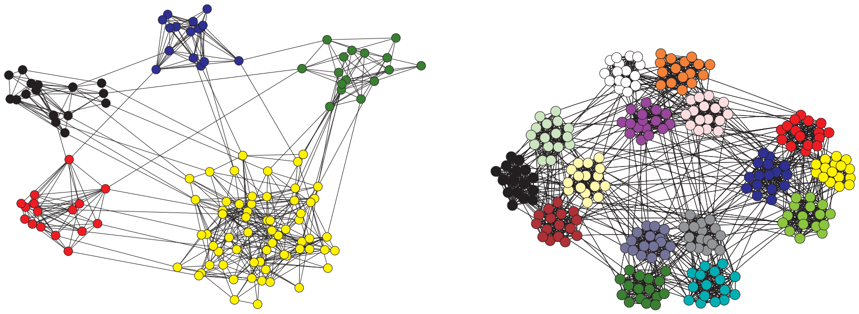


Figure 2.11. Left: Example of a computer generated network with communities of different sizes. In this case we can control the size, the internal cohesion and the external cohesion of each community, see (Danon et al., 2006) for more details. Right: Example of a computer generated network with hierarchical community structure. Each node has a high probability to create a link to nodes of the same internal cluster (z_{in1}), a lower probability to be linked to nodes of the external cluster (z_{in2}) and almost no probability to be linked with nodes of the rest of the network (z_{out}). See (Arenas et al., 2006b) for more details.

Ref.	Label	Order
(Eckmann and Moses, 2002)	EM	$O(m\langle k^2\rangle)$
(Zhou and Lipowsky, 2005)	ZL	$O(n^3)$
(Latapy and Pons, 2004)	LP	$O(n^3)$
(Newman, 2004b)	NF	$O(n\log^2 n)$
(Newman and Girvan, 2004)	NG	$O(m^2n)$
(Girvan and Newman, 2002)	GN	$O(n^2m)$
(Guimerà et al., 2004)	SA	parameter dependent
(Fortunato et al., 2004)	FLM	$O(n^4)$
(Radicchi et al., 2004)	RCCLP	$O(n^2)$
(Donetti and Muñoz, 2004, Donetti and Muñoz, 2005)	DM/DMN	$O(n^3)$
(Bagrow and Bollt, 2005)	BB	$O(n^3)$
(Capocci et al., 2004)	CSCC	$O(n^2)$
(Wu and Huberman, 2004)	WH	$O(n+m)$
(Palla et al., 2005)	PK	$O(\exp(n))$
(Reichardt and Bornholdt, 2004)	RB	parameter dependent
(Newman and Leicht, 2007)	NS	$O(n^2\log n)$
(Danon et al., 2006)	DDA	$O(n\log^2 n)$
(Pujol et al., 2006)	PBD	$O(n\log^2 n)$

Table 2.1. Table summarizing how the computational cost of different approaches scales with number of nodes n, number of links m and average degree $\langle k\rangle$. The labels shown here are used in figures 2.9.

3.3 Which algorithm should we use?

One has to take many factors into account when choosing an algorithm to use. The above comparison ought to give the reader an idea as to which algorithm is most appropriate for a given problem. In many cases, a compromise must be reached between accuracy and running time, especially for larger networks. To clarify this further, here are a few examples of real networks, and our suggestion for the appropriate community identification algorithm.

Say we want to analyze a relatively small network, for example the metabolic network of the worm *Caenorhabditis elegans*, which has 453 nodes. Since the network is small, and current desktop computer technology is reasonably fast, the speed of the algorithm should pose no restriction, and one is free to chose the slower, more accurate methods. In this case the Simulated Annealing (SA) method would be the most appropriate choice, since it gives the most accurate partitions, especially if the system is allowed to cool slowly (see (Guimerà et al., 2004, Massen and Doye, 2005, Guimerà and Amaral, 2005a) for more details).

Larger networks, with the number of nodes in the order of 10^5 become intractable with the most accurate methods. For example, when attempting to study the community structure of the actor collaboration network with 374511 nodes, we estimate that the SA would take a few months of uninterrupted computation. However, a reasonable implementation of the fast algorithm would be able to perform this analysis in just a few hours (Clauset et al., 2004, Pujol et al., 2006), making it the appropriate choice, even if their accuracy is not the best.

Finally, let us consider an intermediate sized network such as the Pretty Good Privacy (PGP) web of trust social network (Guardiola et al., 2002), containing 10680 nodes. Although the SA algorithm would run in a reasonable time, it may be a better choice to compromise and employ a faster running algorithm. We leave this choice to the preferences of the researcher, since all the methods presented in this chapter can perform reasonably well (with more or less accuracy).

4. Summary

In this chapter we have presented the problem of community identification in complex networks, and we have given a brief overview and comparison of the modern approaches to detect the communities. A large amount of knowledge has been collected in the field, and real progress has been made, both in the identification of communities and their characterization. However, some questions do remain open in the community detection problem, and it is these that we would suggest for further study.

One of the main problems that is actually being discussed is the validity of the modularity as the appropriate measure for quantifying the community structure (Arenas and Díaz-Guilera, 2007). The work of Fortunato and Barthelemy showed the limitations of this measure to uncover certain well-defined communities, opening the door for other possible structure measurements. This work is also related with the problem of the detection of community structure at different mesoscales. As introduced previously, it has been observed that the community structure can be analyzed at different levels, obtaining different coarse-grained views of the same network. The structures obtained at different scales do not necessarily need to be hierarchical. The methodology and the definition of this scales are still an open question.

An additional issue that is actually discussed is the computational cost needed to uncover some community structures with the presented methods. The fastest algorithm runs in linear time, but this particular method needs a priori knowledge of the number of expected communities, and assumes that all communities are of similar size (Wu and Huberman, 2004). And if we do not know the number of communities a priori, the cost of the best method scales as $O(n \log^2 n)$ with network size. While this makes the analysis of extremely large networks feasible, this algorithm does not guarantee that the partition found is the best possible one. Other algorithms which are more computationally expensive have other merits, such as accuracy or the ability to identify overlapping communities. So, when choosing a method one must consider carefully the context of its use. Ideally, one would like to have a method which guarantees accuracy and is fast at the same time, but finding such a method is still a challenging problem.

Chapter 3

DETECTING COMMUNITY STRUCTURE USING EXTREMAL OPTIMIZATION

In the previous chapter we have introduced the problem of detecting the community structure as one of the most challenging open problems within the subject of complex networks. As we have seen, the problem has been tackled from several perspectives, but the results are still far to be optimal: the most accurate methods are usually not scalable, and the fastest methods usually do not find the expected communities. The purpose of this chapter is to introduce a novel method to detect the communities based on the maximization of the modularity Q measure. We propose a fast, scalable algorithm that searches highest possible value of Q using local information.

It has been proved that the search for the optimal (largest) modularity value is a NP-hard problem due to the fact that the space of possible partitions grows faster than any power of the system size (Brandes et al., 2007). For this reason, a heuristic search strategy is mandatory to restrict the space of configurations while preserving the optimization goal. Indeed, it is possible to relate the current optimization problem for Q with classical problems in statistical physics, e.g. the spin glass problem of finding the ground state energy (Sherrington and Kirkpatrick, 1975), where algorithms inspired in natural optimization processes as simulated annealing (Kirkpatrick et al., 1983) and genetic algorithms (Goldberg, 1989) have been successfully used.

The heuristic search proposed in this chapter is based on the Extremal Optimization (EO) algorithm introduced by Boettcher and Percus (Boettcher and Percus, 2001a, Boettcher and Percus, 2001b). This algorithm is inspired in turn in the evolution model of Bak-Sneppen (Bak and Sneppen, 1993), and basically operates optimizing a global variable by improving extremal local variables that involve co-evolutionary avalanches. The performance of EO algorithms have been shown to overcome the efficiency of classical simulated annealing and genetic algorithms providing competitive accuracy but using less computational time (Boettcher and Percus, 2000).

The chapter is organized as follows. In the first section we describe the idea behind the extremal optimization heuristics, and how we have applied it to identify communities. Next, we introduce different theoretical and algorithmic improvements that can be applied to increase the accuracy and efficiency of community detection methods based in general. In particular, we show how to apply this refinements to the EO, obtaining even higher values of Q. The last part of the chapter presents a detailed benchmark realized with the EO method. We analyze a large subset of real networks and we show the maximum modularity obtained for each, providing a reference where other algorithms can compare their accuracy.

1. Extremal Optimization Algorithm

The problem of finding an optimal solution for NP-hard problems has received a lot of attention from computational complexity theory. A paradigmatic case of an NP-hard problem is the traveling salesman problem (TSP), which can be formulated as “Given a number of cities and the costs of traveling from any city to any other city, what is the cheapest round-trip route that visits each city exactly once and then returns to the starting city?” (Lawler et al., 1985). Exact solutions can be found for small TSP sizes using exhaustive analysis of the different combinatorial possibilities, but as the system size grows, the number of possibilities to explore is too large to face with the actual computational resources. In this case one should use algorithms that can provide very good solutions, but which could not be proved to be optimal.

The TSP has received a large number of heuristic algorithms that have been specifically designed to find an optimal solution for this problem. However, many physical problems do not have a specialized heuristic procedure to find the solution. For this group of generic problems, the scientific community has developed some general-purpose optimization approaches based on stochastic procedures. Probably the most famous is the Simulated Annealing (SA) algorithm introduced by Kirkpatrick (Kirkpatrick et al., 1983), which is inspired in the behavior of physical systems in thermal equilibrium. Simulated annealing works by taking an initial state of the system, and then trying to improve the system performing small changes, accepting them if they improve the overall status. These changes can drive the system sometimes to better and sometimes to worse optimal states, which are governed by the laws of equilibrium statistical physics. If the energy differences between this local optimum are large (i.e. they are surrounded by high energy barriers), the search algorithm can get trapped in this local optimum without the possibility of continue improving. Therefore, this type of local search methods usually need some kind of mechanism that can help the system to hop between local optimum (see figure 3.1).

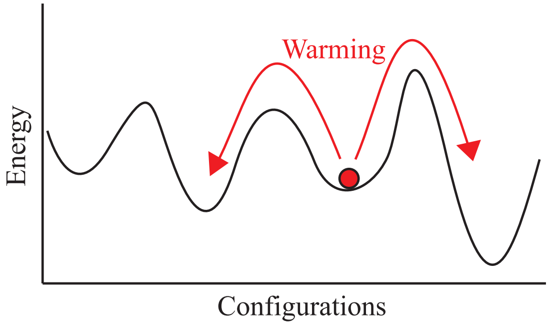


Figure 3.1. Example of the simulated annealing heuristic search process in a one-dimensional configuration space. Each configuration has associated a given energy value. With the simulated annealing we look for the configuration that minimizes the energy of the system. The red ball represents one run of the simulated annealing that is trapped in a local minima. When we increase the temperature of the system, we allow the ball to jump between local minima, potentially making any configuration accessible.

In Simulated Annealing this is done with a temperature parameter that allows to heating or cooling the system. Therefore, we can explore a wider range of possibilities which can be closer to the best optimal result.

Other examples of generic purpose heuristic algorithms that have been used in statistical physics are the Genetic Algorithms (Holland, 1975) or the Tabu Search (Glover, 1986). Within this context, Boettcher and Percus introduced a new local heuristics known as Extremal Optimization that is based on the observation of optimization processes in natural systems (Boettcher and Percus, 2000). The inspiration comes from the natural selection process: the survival of one species depends on the overall adaptation of their population, and to maintain or improve this overall status, the less adapted elements should be discarded. Bak and Sneppen modeled this evolution in an ecological model of interacting species that co-evolve through chain reaction called avalanches (Bak and Sneppen, 1993). The idea is very intuitive. Each species is characterized by a fitness value which measures its adaptation to the environment. The species with the worst fitness (i.e. the less adapted) is selected and it is assigned with a new random fitness value. But the changes on the fitness of one species impacts the fitness of the interrelated species, provoking an avalanche of changes that rearrange the fitness of a large number of elements. After a certain number of steps, the system reaches a punctuated equilibrium (Eldredge and Gould, 1972), with states that remain stable for long time, broken by periods of burstiness where the system evolves very quickly into another metastable state. The most interesting point is that the evolutionary process is performed without any external forces governing the dynamics (like the temperature in SA). This type of behavior resembles de phenomenon of Self-organized criticality (SOC) in statistical physics (Bak et al., 1987), and is considered one of the causes of the emergence of complexity in natural systems (Bak, 1996).

The Extremal Optimization can be considered as a generalization of the Bak-Sneppen model. Using EO we can try to find a near-optimal solution to any NP-problem in a reasonable time. The analyzed problem should be able to be decomposed in terms of an space of possible configurations, and each configuration should be assigned with a global magnitude (the value that we want to optimize). Like in the BS model, each element of the system is assigned with a fitness value that reflects its participation to this global magnitude. Then, the dynamic process works selecting the worst value and replaces its fitness by a randomly new value. After a certain number of changes, the system evolves into a critical state that gives an optimal configuration possible of its elements that maximizes the global value.

In contrast with the thermal equilibrium dynamics of the SA, behind the EO algorithm there is a mechanism that “drives the system far from equilibrium” (Boettcher and Percus, 2002). In SA we analyze each small modification of the configuration and we accept it according to the Metropolis criteria. However, since in EO the system self-organizes, there is no need to decide if we accept a given change. Instead we accept all the changes of the system, which usually are in form of avalanches, and we measure the configuration when the system has stabilized. This could seem an ineffective random search but, as Boettcher and Percus proved, the persistent elimination of the worst fitness values leads the system into meta-stable sub-optimal solutions that can be better than the ones found by SA. To illustrate the performance of the EO, they applied it to the well-known problem of graph bi-partitioning also introduced in chapter 2. Their results showed that the EO outperforms other heuristic methods such as simulated annealing or genetic algorithms, obtaining better results and consuming less time and computational resources (Boettcher and Percus, 2000).

1.1 Optimizing the modularity

The community detection problem can be viewed as a graph multi-partitioning problem where, instead of minimizing the cut size, the main goal is to find the configuration that maximizes the modularity. Remember that modularity is considered the ’de facto’ quantitative measurement for the community structure, which was originally formulated by Newman as:

$$Q = \sum_r (e_r - a_r^2) \tag{3.1}$$

where e_r refers to the fraction of internal links in community r and a_r refers to the total number of links that have at least one node inside community r. Since we are interested in applying the EO algorithm to maximize the value of Q, we need to reformulate equation 3.1 to reflect the individual contribution of the nodes, while maintaining Newman's original idea: the comparison between the internal link connectivity against the expected connectivity in a random network:

$$Q = \frac{1}{2L} \sum_i \sum_j \left(k_{ij} - \frac{k_i k_j}{2L} \right) \delta(C_i, C_j)\,, \tag{3.2}$$

where k_{ij} is 1 if there is a link between i and j and 0 otherwise, k_i is the degree of node i, the Kronecker delta function $\delta(C_i, C_j)$ takes the values, 1 if nodes i and j are into the same community, 0 otherwise, and the number of links $L = \frac{1}{2} \sum k_i$. From this equation we can easily extract the contribution of individual nodes i to the summation,

$$Q = \frac{1}{2L} \sum_i q_i \tag{3.3}$$

being

$$q_i = \sum_j \left(k_{ij} - \frac{k_i k_j}{2L} \right) \delta(C_i, C_j) \tag{3.4}$$

For simplification purposes, we rewrite equation 3.4 as the modularity of the node i belonging to the community r:

$$q_i^r = k\mathrm{int}_i - k_i a_r(i) \tag{3.5}$$

where $k\mathrm{int}_i$ is the number of links that a node i has with the nodes belonging to the same community r where i belongs, $k\mathrm{int}_i = \sum_j k_{ij} \delta(C_i, C_j)$, and $a_r(i)$ is the fraction of links that have one node in community r, $a_r(i) = \sum_j \frac{k_j}{2L} \delta(C_j, r)$.

Equation 3.5 provides a measure that depends on the node degree, and its normalization involve all the links in the network after summation. Re-scaling the local variable q_i^r by the degree of node i we obtain a proper definition for the contribution of node i to the community r relative to its own degree, and normalized in the interval [-1,1].

$$\lambda_i^r = \frac{q_i^r}{k_i} = \frac{k\mathrm{int}_i}{k_i} - a_r(i) \tag{3.6}$$

keeping in mind this definition of λ_i^r we can compare the relative contribution of individual nodes to the community structure. We will consider λ_i^r as the local variable involved in the extremal optimization process that characterize an individual node, from now on we will refer to λ_i^r as the fitness of node i using the common jargon in extremal optimization problems.

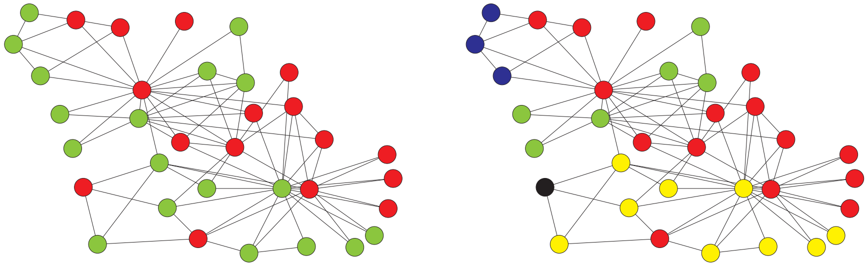


Figure 3.2. Left: Random initialization of the Zachary network into two partitions, red and green. Right: Here we identify five different communities by looking at the connected components in each partition. Each color defines a different community.

Once we have determined the metric that we will use to measure each configuration of the partition space, the next step towards the adaptation of the EO algorithm is the definition of the dynamic process that should self-organize the nodes into communities. The heuristic search we propose to find the optimal modularity evolves as follows:

- Initially, we split the nodes of the whole graph in two random partitions having the same number of nodes each one. This splitting creates an initial communities division, where communities are understood as connected components in each partition.

- At each time step, the system self-organizes by moving the node with the lowest fitness (extremal) from one partition to the other. In principle, each movement implies the recalculation of the fitness of many nodes because the right hand side of equation 3.6 involves the pseudo-global magnitude $a_r(i)$.

- The process is repeated until an “optimal state” with a maximum value of Q is reached. After that, we delete all the links between both partitions and proceed recursively with every resultant connected component. The process finishes when the modularity Q can not be improved[1].

Note that this process is not a bi-partitioning of the graph because: the number of nodes in each partition is dependent on the evolution process and not restricted to be the same at the end of the process; and more importantly, each partition could contain different connected components (communities) that when the partitions are disconnected result in several subgraphs.

[1] The value of Q always refers to the whole network i.e. is the sum over all the communities. At a certain moment more subdivisions into communities will necessarily decrease Q because the limit of decomposition is a community per node whose value of Q is negative.

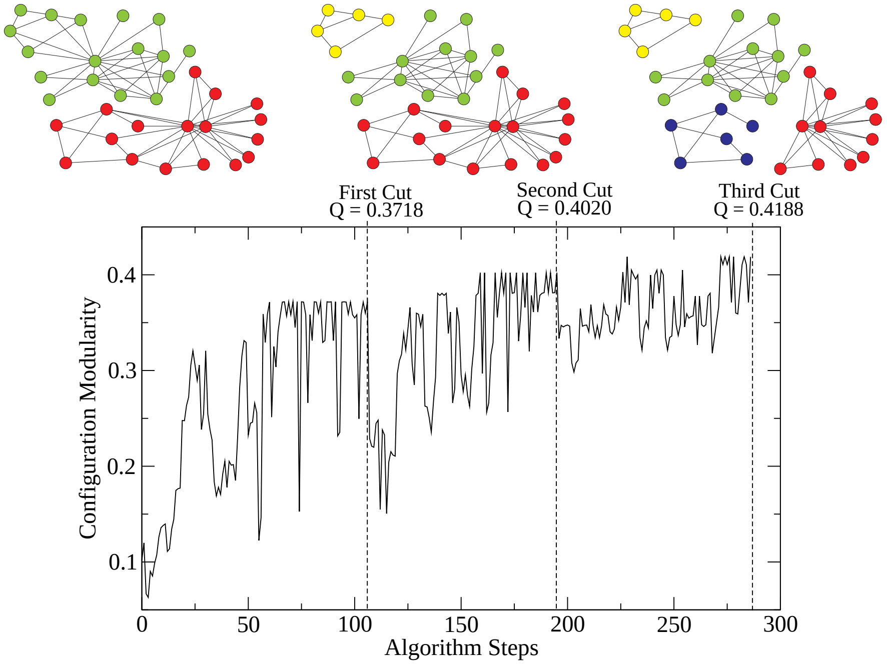
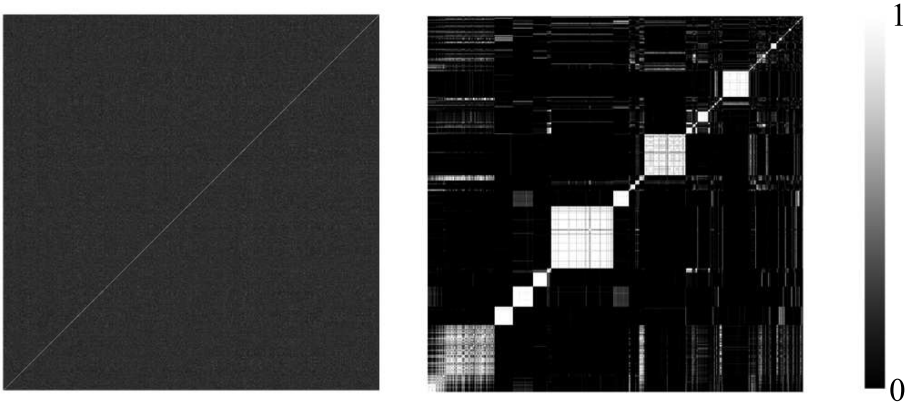


Figure 3.3. Top: Network after edge removal at each recursive cut. Bottom: Evolution of the Q value in the at each step of the adaptation process. Separation bars indicate recursive divisions of the graph performed at maximum Q.

Let us illustrate the above mentioned heuristics in a simple case. We will apply it to the well-know Zachary karate club network (Zachary, 1977). Initially we split the nodes in two random partitions (see figure 3.2 left). Note that the number of initial communities (connected components in each partition) in this case is five (see figure 3.2 right). After that, the self-organization process starts: the node with the "worst fitness" is selected and moved from its partition to the other partition, this movement provokes an avalanche of changes in the fitness of the rest of nodes. We calculate the new value for the modularity Q, and again repeat the process until no changes could improve it (see figure 3.3).

The application of the algorithm to the Zachary network provides the optimal modularity value after three recursive iterations. The network is decomposed in four communities and the value for the modularity is 0.4188, greater than the value 0.381 reported by Newman (Newman, 2004b), the value 0.406 reported by Reichardt et al. (Reichardt and Bornholdt, 2004) and the value 0.412 reported by Donetti et al. (Donetti and Muñoz, 2004) using different optimization methods presented in the previous chapter.



Random network Mail network

Figure 3.4. Fraction of nodes classified in the same partition over 100 realizations of the algorithm. The color of the position (i,j) corresponds to the fraction of times that nodes i and j belong to the same partition.

1.2 Implementation details

The EO approach presented has several technical implementation details that are relevant for our purposes. In the first place, as we introduced previously the main drawback of this type of local heuristics is that the changes produced at each step are usually small and can lead the system to sub-optimal configurations. In the original EO algorithm, the node selected is always the node with the worst λ_i value. This is a deterministic and fast way to solve the problem, but the final result strongly depends on the initialization and there is no possibility to escape from local maxima. Instead, we use a probabilistic selection called τ-EO (Boettcher and Percus, 2001b), in which the nodes are ranked according to their fitness values, and then the node of rank r is selected according to the following probability distribution:

$$P(r) \propto r^{-\tau} \tag{3.7}$$

This solution is less sensitive to different initializations and allows to escape from local maxima. The exponent τ has been tuned around the optimal values obtained for random networks of size N that approach the scaling $\tau \sim 1 + 1/ln(N)$ (Boettcher and Percus, 2001b). The use of this technique also implies the determination of the number of self-organization steps αN needed to decide that the maximum value has little chance to be improved. In practice, we keep track at each step of the last maximum value obtained for Q, if this maximum is not improved in αN steps we stop the search. Usually α is empirically determined balancing accuracy and efficiency in the algorithm, we use $\alpha = 1$ allowing as many steps as nodes to improve the current maximum value of Q.

A second technical detail to consider is the speed of the algorithm. Since one of the objectives is to provide an algorithm as fast as possible, it is important to reduce as maximum as possible how the cost scales with the system size. The computational cost involved in the whole process is $O(N^2 lnN)$ where the $NlnN$ term is the cost associate to the ranking process, however it can be substantially reduced using heap data structures (Aho et al., 1983) for the ranking selection process up to $O(N)$. The total cost of the algorithm can then be improved up to $O(N^2)$. The analysis of a network of 10^5 nodes only takes a few minutes in a standard computer, and network of 10^6 nodes can take up to a day.

And finally, another interesting technical detail that one should care about is the robustness of the algorithm, defined as the capability of finding the same configuration in different runs. Note that since the core of the presented algorithm is stochastic, different runs could yield in principle different partitions. We have performed 100 runs of the algorithm for the e-mail network and for a random network with the same number of links and nodes to check the consistency of the proposed method. In figure 3.4 we present the results of the fraction of times a couple of nodes are classified in the same partition. The same community structure is clearly revealed for the e-mail network while for the random network this structure is inexistent.

1.3 Testing the EO algorithm performance

To test the performance of the algorithm we have used the computer-generated graphs with a known community structure presented in the previous chapter: a network of 128 nodes with 4 communities of 32 nodes, where the nodes have an average degree of 16 and we control the number of internal and external links (Girvan and Newman, 2002). We generate several graphs using z_{out} values between 0 and 10, and we compare the results of our algorithm with those obtained using the heuristics proposed in (Girvan and Newman, 2002) and in (Guimerà et al., 2004). This comparative shows the capabilities of each algorithm identifying the communities when these are more fuzzy inside the whole network.

Using the Girvan-Newman algorithm, which has been the reference algorithm for community identification, the communities are well detected until values of $z_{out} = 6$. In contrast, our algorithm detects the communities up to $z_{out} = 8$, where the community structure still persist but is much more difficult to reveal, see figure 3.5. In this particular case 50% of the links are within the community and 50% are links with nodes outside the community. This result that could seem contradictory is not. Note that the 50% of links with nodes outside the community are equally distributed among the rest of com-

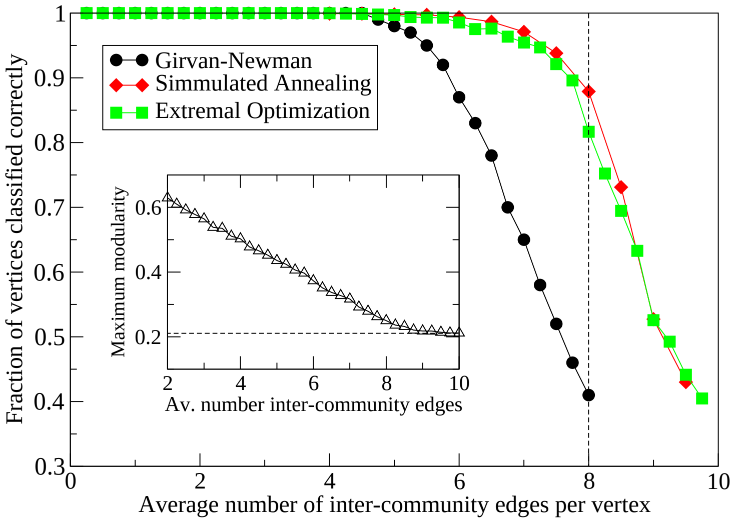


Figure 3.5. Fraction of nodes correctly classified using computer-generated graphs described in text. Each point is an average over 100 different networks. Inset: Average of the maximum modularity obtained in each case.

munities, and then its contribution to the definition of community is deprived by the number of communities in the rest of the network, in our case three. For this reason it is expected to find community structure even in these cases. However, for values higher than 8, the average maximum modularity rapidly approach the limit $Q = 0.208$ (see inset figure 3.5), the expected modularity for a random network with the same number of links and nodes, as it has been shown in (Guimerà et al., 2004).

We also compare the accuracy of the EO algorithm against the simulated annealing algorithm of by Guimera *et al.*, since in the previous chapter we have shown that it is the most accurate algorithm that has been published in the literature. In this case we observe that the EO finds similar values of modularity than the SA. It seems that the SA still performs a little bit better, probably because it can explore a wider space of configurations. But the EO can find these values in less computational time, achieving the two goals that we have presented at the beginning of the chapter, speed and accuracy, and therefore providing a very good alternative to detect community structure to the existent methods.

2. Detecting weighted and directed communities

A large number of real networks are originally weighted and directed. However, the initial complex networks theories have been centered only in the analysis and modeling of unweighted and undirected versions. So when we want to perform a detailed analysis of many real networks, we need first to convert them into their undirected and unweighted versions, usually throwing away useful information which may help us understand the network structure more accurately.

To solve this problem, some authors have started to cope with the statistical analysis of weighted and directed networks, establishing the bases for these scenarios. The first important publications have shown that the inclusion of weights can change substantially the description of real networks (Barrat et al., 2004a, Barrat et al., 2004b). A few attempts have also been proposed to uncover the community structure using the weight information of the links in (Newman, 2004a, Palla et al., 2007) or the directional information in (Farkas et al., 2007).

The original version of the Extremal Optimization algorithm is only capable of detecting communities in undirected and unweighted networks. Here we present a generalized method that takes into account the link directions and weights. The adaptation process is very simple; we maintain the core of the EO algorithm and we need only to redefine the modularity measurement to include the extra information of the links. Preserving its semantics in terms of probability, the definition of modularity can be rewritten as:

$$Q' = \frac{1}{2W} \sum_i \sum_j \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2W} \right) \delta(C_i, C_j) \,, \tag{3.8}$$

where w_i^{out} and w_j^{in} are respectively the output and input strengths of nodes i and j

$$w_i^{\text{out}} = \sum_j w_{ij} \,, \tag{3.9}$$

$$w_j^{\text{in}} = \sum_i w_{ij} \,, \tag{3.10}$$

and the total strength can be computed now as the sum of outlinks or the sum of inlinks

$$2W = \sum_i w_i^{\text{out}} = \sum_j w_j^{\text{in}} = \sum_i \sum_j w_{ij} \,. \tag{3.11}$$

Note that when the network is undirected, the input and output strengths are equal ($w_i = w_i{}^{\text{out}} = w_i{}^{\text{in}}$), and we obtain the modularity as a function only of the strength. Furthermore, if the network is unweighted and undirected, w_i represents the degree of the i-th node, i.e. the number of edges attached to it, and W is the total number of links of the network.

Once we have redefined the modularity in terms of weighted and directed links, now we can define the value of the local contribution of the nodes to the weighted modularity as before,

$$Q' = \frac{1}{2W} \sum_i q'_i \tag{3.12}$$

being

$$q'_i = \sum_j \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2W} \right) \delta(C_i, C_j) \tag{3.13}$$

However, the final step to obtain the fitness of each node is not straightforward. In this case, due to the fact that we are using the directed modularity, we have two choices as the fitness of the nodes: we can use the information of the inlinks or the information of the outlinks. For instance, if we choose the contribution of the nodes to the modularity in form of incoming links, the definition of the fitness for node i belonging to community r is

$$\lambda_i^r = \frac{w\text{int}_i^{\text{out}}}{w_i^{\text{out}}} - a_r^{\text{in}}(i) \tag{3.14}$$

where $w\text{int}_i^{\text{out}}$ is the sum of the weights of the outgoing links from node i to other nodes belonging to community r, $w\text{int}_i^{\text{out}} = \sum_j w_{ij}\delta(C_i, C_j)$, and $a_r^{\text{in}}(i)$ is the fraction of the weights of the links that have its destination in community r, $a_r^{\text{in}}(i) = \sum_j \frac{w_j^{\text{in}}}{2W}\delta(C_j, r)$. And if we exchange *out* by in and viceversa in equation 3.14, we obtain a second definition of fitness according to the outgoing links.

$$\lambda_i^r = \frac{w\text{int}_i^{\text{in}}}{w_i^{\text{in}}} - a_r^{\text{out}}(i) \tag{3.15}$$

Which one should we choose as the fitness? The answer is that both are equally valid since when we sum the local contribution of the nodes we obtain the same global value for the modularity. The only difference between using this two fitness is that we could obtain a distinct node ranking, provoking that the heuristic search paths through the configuration space will also be different.

To analyze the performance of this extension of the original EO, we have tested it on a set of computer generated networks that include links with weight and direction. These networks are created using the same methodology described in the previous chapter (128 nodes network, divided into four groups of 32 nodes with an average degree of 16), but now each link is assigned randomly with one direction. Since we want to study the effect of weights, we fix the average number of internal/external links and we assign a weight $w >= 1$

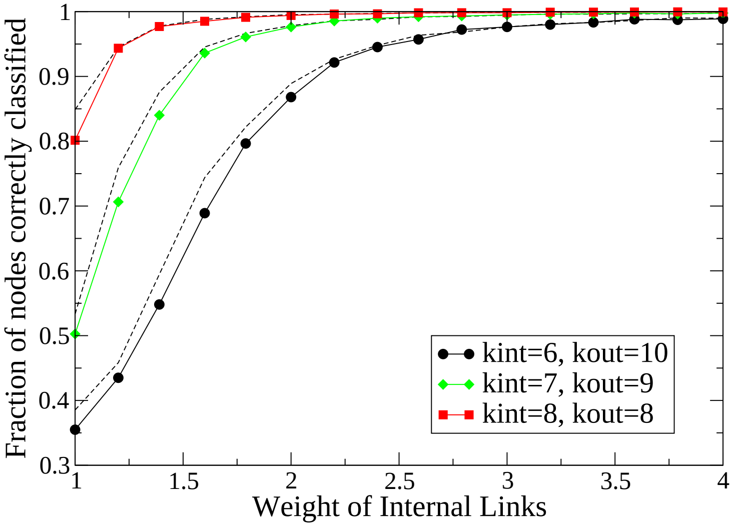


Figure 3.6. Fraction of nodes correctly classified using computer-generated graphs with directed and weighted links. We fix the number of links, the weight of the external links to 1, and we increase the weight of the internal links. Dotted lines represent the accuracy of the detection process when we remove the direction of the links.

to the edges inside each community while we keep a fixed weight 1 for those edges that lie between communities. Then we evaluate the fraction of vertices classified correctly as a function of the internal weight, w. As figure 3.6 shows, introducing weight into the links provides extra information that allows the algorithm to discover the 4 communities again, even if there are more links to outside than to inside the community. We also observe that values of w needed to recover the communities are very small, when the weight of the internal links is twice the weight of the internal, the algorithm classifies essentially all vertices correctly in the three presented cases.

We also compare the results against undirected version of the same networks. Note that, for lower values of w the undirected detection seems to find better configurations, but when the communities are well-defined there is almost no difference between them. This is also an expected result since, as we explain in appendix A, the differences between the directed and undirected versions of modularity of the same network are usually very small.

3. Increasing the efficiency and accuracy of community detection algorithms

After the design of the EO, we have focused on the design of alternative methods to increase the values of the maximum modularity and to reduce even more the time need to find these values. The methods are not specific for the

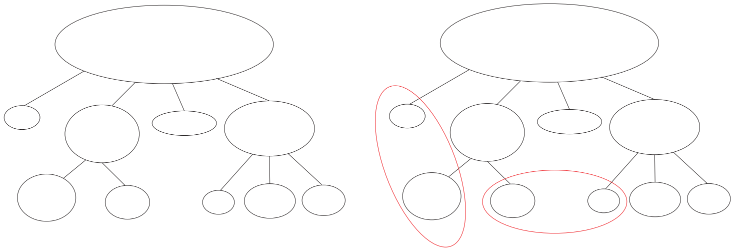


Figure 3.7. Left: Hierarchical tree that represents the recursive process behind the EO algorithm. The final partitions are defined by the leaves of the tree. Right: Combining the EO with Newman's fast algorithm we can explore other possible configurations that can still increase the modularity.

presented algorithm, and can be applied to any algorithms based on modularity optimization.

3.1 Improving recursive algorithms

We begin presenting two different ways to improve the accuracy of the recursive algorithms without incrementing their computational cost. These improvements try to solve two typical problems associated with the recursive heuristic searches, independently if the method is divisive or agglomerative. Other methods that identify all the communities at once, such as Simulated Annealing, explore a larger space of configurations and therefore do not not suffer from this two problems.

The first problem is in terms of the space of configurations that is visited during the recursive analysis. Every time we separate (or group) the network into two or more partitions, we eliminate the possibility of going backwards and explore other configurations. In the particular case of the Extremal Optimization algorithm, this problem is reflected in the large number of communities that we find in comparison with other algorithms that obtain similar modularities but a smaller number of communities (Pujol et al., 2006). This is due to the fact that border nodes of small communities usually have higher values of local modularity than the border nodes of larger communities, and therefore these nodes are rarely selected to change from one partition to another. When the system cuts the two partitions into several groups a large number of small communities are isolated, losing the chance to integrate with the larger ones. As the system size growths, the probability of having small communities also increases.

One easy way to solve this problem is to use a combination of an agglomerative and a divisive algorithm. This combination provides a deeper exploration of the configuration space, since allows the analysis of a new group of potential partitions. Figure 3.7 illustrates how the solution works. First we let the recursive algorithm to divide the network into several groups. Then we perform an agglomerative process to try to re-join some communities that have been separated during the recursive process. Since we work with the output of the first algorithm, there are only a few communities that we can try to merge, so the second algorithm is extremely fast.

In our particular case, we have combined the EO algorithm with Newman's fast algorithm proposed in (Newman, 2004b). The idea of the fast algorithm has been described in the previous chapter: we compute the increment of the modularity obtained by joining each possible pair of communities detected by the EO, we select the highest increment and, if the increment is positive, we merge the two communities. We repeat this process recursively until we can not increase the modularity anymore. As we have explained the process is very fast, since the algorithm runs in almost linear time for sparse networks, $O(n \log n)$, and the number of communities to agglomerate is very small. When we apply this method to the communities found by the EO, we observe that we almost do not increase the final value of the modularity, but instead the number of communities detected is reduced drastically (see table 4.1 in the last section).

The second problem is also an artifact of the recursive mechanism. It can happen that during the initial splits (joins) of the network we obtain an intermediate configuration that temporarily has the best modularity possible. However, after performing recursive splits (joins) it turns out that the intermediate cut that we have performed does not let us to reach the final maximum modularity. We will illustrate this problem using the already mentioned Zachary network.

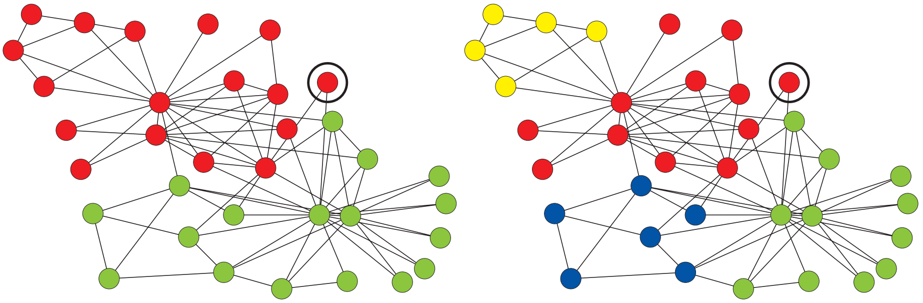


Figure 3.8. Example of a node misclassification problem related with recursive algorithms that optimize the modularity. First we analyze the network and we obtain the two partition depicted in the right picture. The circled node is classified in the red partition since it provides higher modularity than if it is classified in the green. We cut the network into this two groups and we perform recursively the analysis of the two partitions, obtaining the final four partitions shown in the left figure. In this case we cannot obtain the maximum modularity, since the circled node should be in the green partition to obtain the maximum, but we cannot explore this configuration due to the limitations of the recursive process.

When performing the initial analysis, we observe that the circled node is classified in the red partition, as seen in figure 3.8 left. This configuration gives the maximum possible modularity, since if we change the marked node to the other side the modularity decreases from 0.371794 to 0.371466. Therefore, we consider this as the best split, we cut the network and then we apply the recursive procedure, obtaining the four partitions observed in figure 3.8 right. In this case the final modularity is 0.418803. However, this is not the maximum possible modularity that we can obtain with four communities. If we move the marked node to the green partition we can increase the modularity to 0.419790, obtaining the configuration with the highest modularity known for the Zachary network.

In our case we have solved this problem using a final bootstrapping of the modularity, similar to the refinement mechanism introduced by Newman in (Newman, 2006a). In this final step, we let all the nodes to move to other partitions and observe the changes in the modularity, similar to the process explained before in the fast algorithm. When we detect that one change improves the final modularity, we accept the movement of the node. The process is repeated once for all the nodes, so we are able to correct some node positions and obtain configurations with even higher modularity values.

The increase of the modularity obtained in this two refinements is usually very small, usually in the order of 10^{-2}. For instance, moving one node that is isolated into a group will only increase the modularity value in $\min_i\{w_i\}/2w$. Therefore, we must take into account that in the search for the best modularity we need to work with high precision. A difference in the second (or even greater) decimal can carry some important structural difference between two configurations.

3.2 Size reduction of the network

Up to now we have presented two different techniques that can improve the results of community detection algorithms. In this case we present a different approach, proposing a method to reduce of the size of the network instead of improving the problems related with the algorithms. Our goal here is to demonstrate that it is possible to reduce the size of complex networks while preserving the value of modularity, independently on the partition under consideration. The systematic use of this reduction allows for a more exhaustive search of the partitions' space that usually ends in improved values of modularity compared to those obtained without using this size reduction. Therefore, we can obtain even more accurate results in less computational time.

One of the most interesting points of this method is that is also independent of the algorithm that is going to be used a posteriori, being specially useful for the algorithms that look for the configuration with the highest modularity. The only imposed constraint to the algorithm is that it should be able to detect the communities of weighted graphs which include self-loop links.

Reducing a graph preserving modularity

Before explaining the different types of reductions that we can carry out, we will introduce the concept of reduced graph. Let G be a weighted complex network of size N, with weights $w_{ij} \geq 0,\ i,j \in \{1,\dots,N\}$. If the network is unweighted, the weights matrix becomes the usual connectivity matrix, with values 1 for connected pairs of nodes, zero otherwise. We will assume that the network may be directed, i.e. represented by a non symmetric weights' matrix.

Any grouping of the N nodes of the complex network G in N' parts may be represented by a surjective function $R : \{1,\dots,N\} \longrightarrow \{1,\dots,N'\}$ which assigns a group index $R_i \equiv R(i)$ to every i-th node in G. The *reduced network* G' in which each of these groups is replaced by a single node may be easily defined in the following way: the weight w'_{rs} between the nodes which represent groups r and s is the sum of all the weights connecting vertices in these groups,

$$w'_{rs} = \sum_i \sum_j w_{ij} \delta(R_i, r) \delta(R_j, s)\,, \quad r,s \in \{1,\dots,N'\} \tag{3.16}$$

where the sums run over all the N nodes of G. For unweighted networks the value of w'_{rs} is just the number of arcs from the first to the second group of nodes. It must be emphasized that a node r of the reduced network G' acquires a *self-loop* if $w'_{rr} \neq 0$, which summarizes the internal connectivity of the nodes of G forming this group.

The input and output strengths of the reduced network G' are

$$w_r'^{\,\text{out}} = \sum_s w'_{rs} = \sum_i \sum_j w_{ij} \delta(R_i, r) \sum_s \delta(R_j, s) = \sum_i w_i^{\text{out}} \delta(R_i, r)\,, \tag{3.17}$$

$$w_s'^{\,\text{in}} = \sum_r w'_{rs} = \sum_j \sum_i w_{ij} \delta(R_j, s) \sum_r \delta(R_i, r) = \sum_j w_j^{\text{in}} \delta(R_j, s)\,, \tag{3.18}$$

and its total strength $2w'$ is equal to the total strength $2w$ of the original network

$$2w' = \sum_r w_r'^{\,\text{out}} = \sum_s w_s'^{\,\text{in}} = \sum_i w_i^{\text{out}} = \sum_j w_j^{\text{in}} = 2w\,. \tag{3.19}$$

One of the properties of the reduced network is the preservation of modularity, i.e. the modularity of any partition of the reduced graph is equal to the modularity of its corresponding partition of the original network. Each node in the reduced network summarizes the information necessary for the calculation of modularity in its self-loop (that accounts for the intraconnectivity of the community) and its arcs (that account for the total strengths with the rest of the network). The proof of this property is available in the Appendix B of this dissertation.

Undirected and directed reductions

The question now is: how can we determine which nodes will belong to the same community in the optimal partition, before this partition is obtained? To answer this questions we need to be able to determine the acquaintance (node j) of node i in its optimal community, in order to group them ($R_i = R_j$) in a single equivalent node with a self-loop, as explained above. If we know that nodes i and j share the same community at maximum modularity, the reduced network will be equivalent to the original one as regards modularity: no information lost, and a smaller size. Taking into account that the sign of the local modularity when a single node i is connected to community r can only be positive if there is a link between i and another node in community r, the only candidates to be the right acquaintance of any node are its neighbors in the network. Here we present two reductions for undirected and two for directed networks based on this idea that does not alter the final modularity. The analytical proof of these reductions has also been included in Appendix B.

In undirected networks, the simplest particular cases are *hairs*, i.e. nodes connected to the network with only one link. Hence, a hair can be analytically grouped with its neighbor k if

$$w_{ii} \leq \frac{w_i^2}{2w}\,, \tag{3.20}$$

producing a self-loop for node k of value

$$w'_{kk} = w_{ii} + 2w_{ik}\,. \tag{3.21}$$

When node i has no self-loop ($w_{ii} = 0$) this condition is always fulfilled, see figure 3.9a.

Another solvable structure in undirected networks is what we call a *triangular hair*, in which two nodes i and j have only one link connecting them, two more links from i and j to a third node k, and possibly self-loops. In this case, if

$$w_{ii} \leq \frac{w_i^2}{2w} \text{ and } w_{jj} \leq \frac{w_j^2}{2w} \tag{3.22}$$

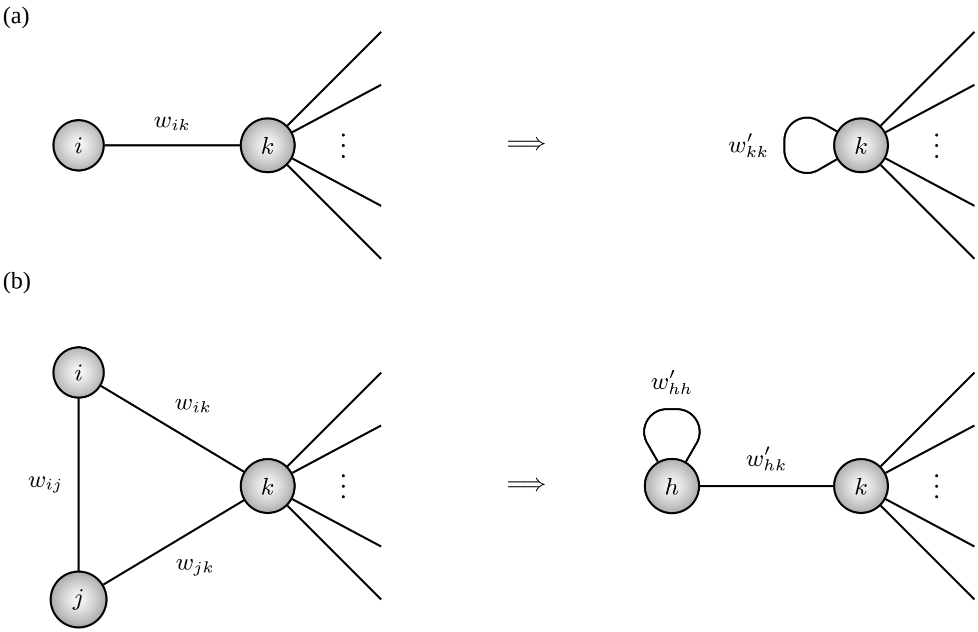


Figure 3.9. Analytic reductions for undirected networks. In (a) example of a *hair* reduction, (b) example of a *triangular hair* reduction (see text for details). The widespread case of unweighted networks, all weights equal to 1, implies that in the reduction (a), $w'_{kk} = 2$, and in the reduction (b), $w'_{hh} = 2$ and $w'_{hk} = 2$.

nodes i and j share the same community in the optimal partition and therefore may be grouped as a single node h. Moreover, the resulting structure becomes a simple hair, which can be grouped with node k if

$$w'_{hh} \leq \frac{w'^2_h}{2w'} \tag{3.23}$$

where

$$\begin{aligned} w'_{hh} &= w_{ii} + 2w_{ij} + w_{jj} \, , \\ w'_{hk} &= w_{ik} + w_{jk} \, , \\ w'_h = w_i + w_j &= w'_{hh} + w'_{hk} \, . \end{aligned} \tag{3.24}$$

In the particular case of nodes i and j without self-loops ($w_{ii} = w_{jj} = 0$), the triangular hair can always be reduced to a single hair with a self-loop $w'_{hh} = 2w_{ij}$, see figure 3.9b.

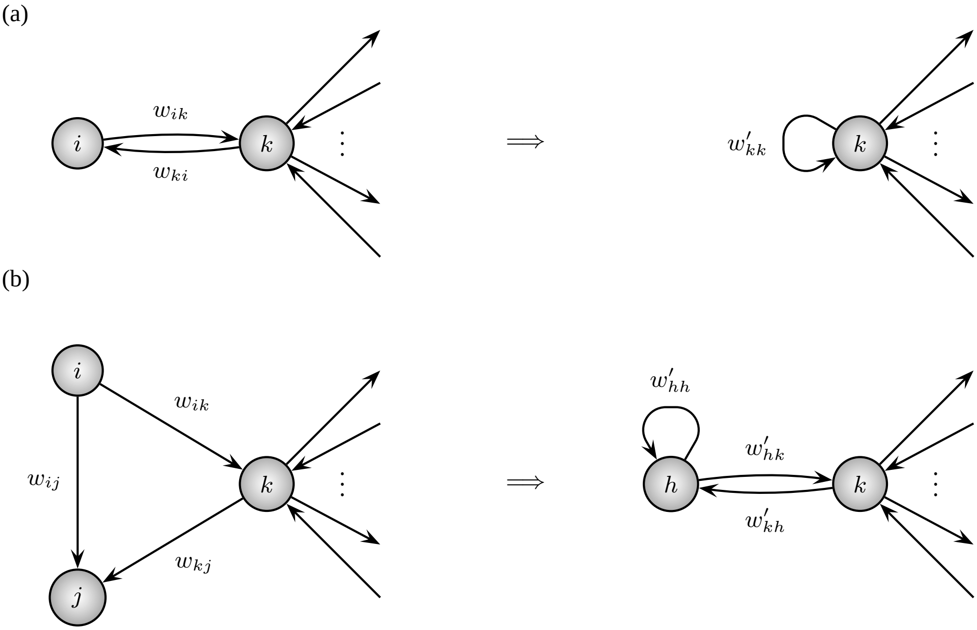


Figure 3.10. Analytic reductions for directed networks. In (a) example of a *hair* reduction, (b) example of a *triangular hair* reduction (see text for details)

In directed networks we can also apply the same reductions. In the case of directed hairs, we can reduce nodes connected only to another node either through an input, an output, or both links. Therefore, it is safe to group them in the same way as undirected hairs if

$$w_{ii} \leq \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \,. \tag{3.25}$$

This condition is always fulfilled if the hair has no self-loop ($w_{ii} = 0$), see figure 3.10a. Whenever the self-loop is present, both input and output links are needed to counterbalance it. The resulting self-loop w'_{kk} of the grouped node has value

$$w'_{kk} = w_{ii} + w_{ik} + w_{ki} \,. \tag{3.26}$$

The case of the triangular hair is more complicated. First we need to define sink nodes as nodes i which are characterized by null output strengths, $w_i^{\text{out}} = 0$, and source nodes, which are defined as nodes with null input strengths, $w_i^{\text{in}} = 0$. Note that sinks and sources cannot have self-loops, since this would be in contradiction with their null output and input strengths respectively. As proved in Appendix C, a triangular hair formed by a source node i and a sink node j behaves exactly as the undirected triangular hair, being possible to group them in a single node h with a self-loop, see figure 3.10b, where

$$\begin{aligned} w'_{hh} &= w_{ij} \,, \\ w'_{hk} &= w_{ik} \,, \\ w'_{kh} &= w_{kj} \,. \end{aligned} \tag{3.27}$$

How much can we reduce a network?

Since the presented reductions can be only applied to very particular cases, a logical question is the amount of reduction that we can obtain applying this method to real networks. Here we provide some rough estimates for the most widespread degree distributions in natural and artificial networks: scale-free and exponential.

For scale-free networks it is usually assumed a $P(k) = \alpha k^{-\gamma}$, with $\gamma \in [2, 3]$ for most of the real scale-free complex networks, as seen in Chapter 1. The normalization condition provides with the value of α. As a first approximation, neglecting the structural cut-off of the network, we can write

$$\alpha \sum_{k=1}^{\infty} k^{-\gamma} = \alpha\zeta(\gamma) = 1 \tag{3.28}$$

where $\zeta(\gamma)$ is the Dirichlet series representation of the Riemman zeta function. For values of $\gamma \in [2, 3]$ we obtain $\alpha \in [1/\zeta(2), 1/\zeta(3)] \approx [0.61, 0.83]$. That means that, roughly speaking, the number of hairs that corresponds to $P(1)$ is about 83% of nodes in a scale-free network with $\gamma = 3$ and 61% when $\gamma = 2$, although this value is slightly reduced when considering the cut-offs of the real distributions.

An equivalent estimate can be conducted for exponential degree distributions of type $P(k) = \alpha e^{-\beta k}$, with $\beta > 0$. In this case, normalization implies that

$$\alpha \sum_{k=1}^{\infty} e^{-\beta k} = \alpha \frac{e^{-\beta}}{1 - e^{-\beta}} = 1 \tag{3.29}$$

and then $\alpha = e^{\beta} - 1$. The percentage of hairs in this case is $P(1) = 1 - e^{-\beta}$, that, for example, for plausible values of $\beta \in [0.5, 1.5]$ provides a reduction between 40% and 77% respectively.

To complement this analytical study, we have also studied the effect of the size reduction applied to real networks. In the following section we will show how the size reduction increases substantially the maximum modularity and the speed of the Extremal Optimization algorithm.

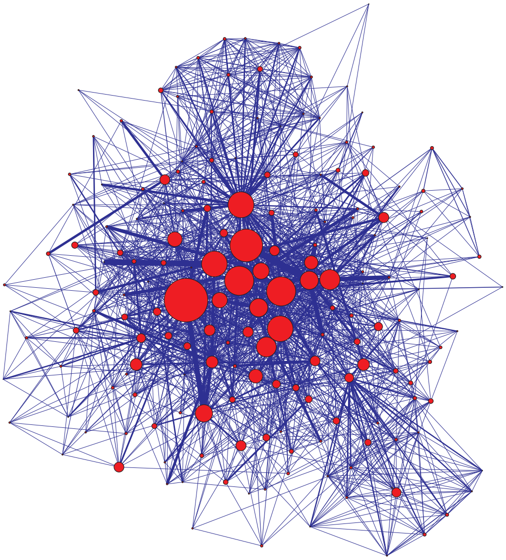


Figure 3.11. Community structure of the FP6 analysis represented as a network. Nodes correspond to communities and links represent collaboration between members of the two connected communities. Diameter of nodes and width of links symbolize community size and number of crossed collaboration.

At the light of these estimates, the size reduction process provides with an interesting technique to confront the analysis of community structure in networks by maximizing modularity. Since the method to detect the single nodes (or even the triangles) can be implemented in O(n), we can obtain with a substantial advantage in computational cost without sacrificing any information. We think that the idea of the exact reduction could be extended to other specific motifs (building blocks) in the network, although its analytical treatment can be further difficult.

4. Uncovering the community structure of real networks

The artificial benchmarks presented before are useful to compare or to test the efficiency of a given algorithm under certain circumstances, but the veritable purpose of the EO algorithm (and the other community detection methods) is to uncover the community structure of real networks, without knowing a priory neither the number nor the size of the communities.

Therefore, since we can not check the results like in the artificial networks, the validity of the obtained configuration depends only on the interpretation that we can perform from the resultant communities. For instance, in social networks the division into communities can be checked by crossing relational data (who is preferably linked to whom) with particular information about each

Community 19
Centre Suisse d'ectronique et Microtechnique
EADS Deutschland Corporate Research Center
Lunds Universitet
Skoda Auto AS
Volkswagen Ag
Robert Bosch Gmbh
Technische Universitat Darmstadt
System Design and Research Association SRL
European Road Transport Telematics Organisation
Audi Aktiengesellschaft
Bayrische Motoren Werke Aktingesellschaft
Bmw Forschung und Technik Gmbh
Seat Centro Tecnico
Volvo Car Corporation
Blaupunkt Gmbh
Delphi Delco Electronics Europe Gmbh
Faurecia Sieges D'Automobile SA
Ibeo Automobile Sensor Gmbh
Siemens Vdo Automotive Sas
Fcs Simulator Systems
Federal Highway Research Institute
Essex County Council
Landeshaupstadt Hannover
Ministry Economics and Transport of Lower Saxony
Laboratory of Lighting Technology. Darmstadt Univ.

Table 3.1. Nodes belonging to one of the resultant communitiess obtained when applying the EO method to the FP6 network. Note that all organizations are related, in some sense, with automobiles. See (Lozano et al., 2006) for a complete analysis of this network.

node. As an illustrative example, we built up a network from a database of research projects of the European 6th Framework Programme, calculated its community structure and analyzed the resulting data by crossing it with information about nationality and organization's type of activity. The community structure obtained is presented in figure 3.11. When analyzing in detail the member of one obtained communities (see table 3.1), we observe that all of them belong to the same activity sector, in this case the automobile market. This social interpretation have also given successful results in other social network community analysis, such as the departmental structure in the communities of the e-mail network of the Universitat Rovira i Virgili (Guimerà et al., 2003) or the racial segregation observed in the community structure of the jazz bands (Gleiser and Danon, 2003).

4.1 Community detection benchmark

In this section we present an exhaustive analysis of the maximum modularity that the extremal optimization method is capable to find. We have performed this analysis in the most used networks in community literature. For each network we have analyzed the maximum modularity and the number of communities found before and after applying the different upgrades presented in this paper. Then we have applied the size reduction to all the presented, networks, and we have repeated the same analysis, but now including the comparison between the percentage of size reduction and the speed-up obtained in the extremal optimization.

The networks analyzed are: the Zachary's karate club network (Zachary, 1977), the Jazz musicians network and the Jazz bands network (Gleiser and Danon, 2003), the e-mail network of the University Rovira i Virgili (Guimerà et al., 2003), the worldwide airports network with data about passenger flights operating in the time period November 1, 2000, to October 31, 2001 compiled by OAG Worldwide (Downers Grove, IL) and analyzed in (Guimerà et al., 2005), the network of users of the PGP algorithm for secure information transactions (Boguñá et al., 2004), the Internet network at the autonomous system (AS) level as it was in 2001 and 2006 reconstructed from BGP tables posted by the University of Oregon Route Views Project, the network of projects involved in the European Union Sixth Framework Programme (also known as FP6) (Lozano et al., 2006), the US airport network collected in 1997, the adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens (Newman, 2006a), the dolphins associations network of the Doubtful Sound community, New Zealand (Lusseau et al., 2003), the network of American football games between Division IA colleges during regular season Fall 2000 (Girvan and Newman, 2002), the collaboration network of contributors to the Spanish statistical physics conference (Fises) (Arenas et al., 2004), the co-appearance network of characters in the novel Les Miserables, (Knuth, 1993), the Western States power grid network of the United States (Watts and Strogatz, 1998), the C. Elegans metabolic network (Jeong et al., 2000), and finally the relations between authors that shared a paper in cond-mat (Newman, 2001a). The results obtained are reported in table 4.1.

We present in table 4.1 the values of modularity for the different networks analyzed up to order 10^{-6}. As we have introduced previously, the numerical resolution of modularity is up to order $\min_i\{w_i\}/2w$, that represents the minimal possible change in the structure of the partitions. It means that every digit in our value of modularity is significant for comparison purposes.

Network	N	Q_1	$\#c_1$	Q_2	$\#c_2$	%r	SU
Zachary	34	0.418803	4	0.419790	4		
Zachary-red	33	0.418803	4	0.419790	4	2,94	1
Zachary-W	34	0.444904	4	0.444904	4		
Zachary-W-red	33	0.444904	4	0.444904	4	2,94	1
Dolphins	62	0.526799	4	0.528519	5		
Dolphins-red	53	0.528519	5	0.528519	5	14,51	1,30
Les Miserables	77	0.551762	7	0.560008	6		
Les Miserables-red	59	0.551762	7	0.560008	6	23,37	1
Word Adjacencies	112	0.303651	6	0.308396	7		
Word Adjacencies-red	102	0.306533	7	0.309154	6	8,92	1,05
Football	115	0.604570	10	0.604570	10		
Football-red	115	0.604570	10	0.604570	10	0	1
Jazz Bands	198	0.444469	3	0.445144	4		
Jazz Bands-red	193	0.444469	3	0.445144	4	2,52	1
US Airports	332	0.360089	9	0.368244	6		
US Airports-red	270	0.359568	9	0.368244	6	18,67	1,17
Celegans Metabolic	453	0.437907	14	0.452021	10		
Celegans Metabolic-red	447	0.437390	11	0.451288	10	1,32	1
Fises	840	0.804918	38	0.827127	22		
Fises-red	722	0.806328	37	0.827352	23	14,04	1,12
E-Mail	1133	0.572024	21	0.580070	10		
E-Mail-red	981	0.574372	16	0.581425	10	13,41	1
Jazz Musics	1265	0.594876	25	0.600561	18		
Jazz Musics-red	1263	0.597452	27	0.600716	18	0,15	1,01
Worldwide Airports-WU	3618	0.642288	146	0.649268	29		
Worldwide Airports-WU-red	2763	0.644834	99	0.649337	29	23,63	1,58
Worldwide Airports-WD	3618	0.643562	116	0.649189	34		
Worldwide Airports-WD-red	2880	0.641834	159	0.649286	30	20,39	1,63
Worldwide Airports-U	3618	0.681868	108	0.706704	25		
Worldwide Airports-U-red	2763	0.682552	89	0.707076	24	23,63	1,68
FP6	3030	0.851265	168	0.877809	48		
FP6-red	3008	0.852480	169	0.878325	47	0,72	1,01
Power Grid	4941	0.896651	132	0.931571	36		
Power Grid-red	3695	0.906516	121	0.933613	39	25,21	2,46
PGP	10680	0.817271	930	0.876883	118		
PGP-red	6277	0.837389	532	0.880244	101	41,22	4,27
AS2001	11174	0.567733	338	0.619048	25		
AS2001-red	7386	0.594702	231	0.628004	31	33,90	2,41
AS2006	22963	0.575363	833	0.645942	49		
AS2006-red	15118	0.611360	428	0.658198	45	34,16	2,38
Condmat	27519	0.617801	2107	0.698032	131		
Condmat-red	24757	0.627627	1632	0.707443	126	10,03	1,12

Table 3.2. Results for the optimal partition obtained using the upgraded EO algorithm for several real networks before and after applying the size reduction. We present the number of nodes, modularity, number of communities (#c), percentage of size reduction (%r) and speed-up (SU) of the algorithm after reduction. We also compare the modularities and the number of communities obtained before using the two presented improvements of the extremal optimization algorithm (1) and after adding those improvements (2).

The first thing that we can notice is that the reduction process allows for a more exhaustive search of the partitions' space as expected. The reductions vary between 20% and 40% in larger networks, which is lower than the predicted in the previous section. The speed-up of the algorithm after reduction gives an indication of the effectiveness of the process. This is also corroborated by an improvement in modularity. Note the special case of the FP6 network, with a reduction of only 0.7%. The lack of reduction is because this network comes from a projection of a highly connected bipartite network, and therefore is mainly composed by groups of highly connected nodes and does not have isolated nodes that can be simplified.

Particularly illustrative is also the analysis of the worldwide airport network. We have constructed different networks from the raw data, the undirected unweighted network previously used in (Guimerà et al., 2005), the undirected weighted network (where the weights reflects the number of passengers using the connection in the period of study), and the most realistic case corresponding to the weighted directed network of the airports connections. These networks allowed us to check our techniques (reduction and optimization algorithm) in all the possible scenarios. Note that the results obtained for the weighted directed and undirected networks in terms of modularity are very close, an explanation about this fact that is ubiquitous in the analysis of directed networks can be found in the Appendix A.

Additionally, if we compare the results of table 4.1 with some of the results in the literature of community detection, we observe that we improve those obtained using Spectral optimization (Newman, 2006b) and simulated annealing (Guimerà and Amaral, 2005b), which have been considered the best up to date. The differences in maximum modularity is up to 15% depending on the network considered.

5. Summary

In this chapter we have presented an extremal optimization based algorithm that optimizes the modularity and allows an accurate identification of community structure in complex networks. The results outperform almost all algorithms existent in the literature. We also have proved that the heuristic process of the EO is very flexible, an for instance can be easily extended to detect community structures in directed and weighted versions.

In second place we have introduced some techniques that can help us to obtain higher modularity values. On one hand, with some algorithmic improvements we can avoid some drawbacks of the recursive process and explore a wider part of the space of possible configurations. On the other hand, we have presented a method to detect the nodes that will be grouped in the maximum modularity configuration, and therefore we can reduce the size of the network preserving the modularity. This reductions are reflected in both increasing the efficiency and the accuracy of the algorithms.

Finally, we have presented an extensive benchmark of the maximum modularity values that the EO method finds in some of the most used networks by the research community. We expect that this benchmark can be used to compare the accuracy of the other modularity based methods that try to unravel the community structure of complex networks.

Chapter 4

SCALING OF FLUCTUATIONS IN TRAFFIC ON COMPLEX NETWORKS

In previous chapters we have introduced a wide group of tools that help us characterize and model the complex topology of many real networks. The study of the topology itself is only the first step towards the understanding of the function of the systems built on networks.

The interest into understanding the dynamical processes that take place on complex networks have emerged recently, and are not as developed as the studies of the structure. The main research lines are trying to uncover the interdependency between the underlying topology and the dynamical processes, answering to questions like: does the dynamical processes affect the shape (and the evolution) of the structure? how does a change on the topology modify the behavior of the functionality?

In this dissertation we are interested in understanding two dynamical properties of a traffic flow, and its relationship with the underlying topology. Traditional approaches to study the traffic have been focused on the study of the long time behavior of a few variables, characterizing phenomena such as the self-similarity (Park and Willinger, 2000). In complex networks, most of the work has been focused in determining the bounds for this flow to become congested (Zhao et al., 2005, Moreno et al., 2003), and how to avoid the congestion to maintain the maximum efficiency of the system (Guimerà et al., 2002b, Echenique et al., 2004, Barthelemy and Flammini, 2006).

A less studied property of the traffic flow in complex networks are the scaling properties of the fluctuations on different nodes of the system. Analyzing how does the standard deviation of each time series change with the value of the mean, we can infer a scaling relationship that gives us more information about the dynamical behavior of the system, answering to questions like: will those elements with larger mean have larger fluctuations as well? what are the reasons behind the differences in the size of the fluctuations?

The study of the existence of this scaling relationship in several time series was first pointed out in the ecology field by Taylor, when he was analyzing the relation between the fluctuations and the average population of a group of species (Taylor, 1961). Since then, several other studies have also uncovered this relationship in a very wide range of systems, such as highways, Internet or the stock market, discovering that each system has a specific relation that helps us to characterize and classify their traffic dynamics (Eisler et al., 2007).

The purpose of this chapter is to show that simple considerations regarding the persistence of packets flowing the network, the limitation of nodes to handle information, and the time window where statistics are performed, account for different scalings of the fluctuations in traffic on complex networks. The chapter is organized as follows: First we introduce the scaling of fluctuations as a large-scale metric to characterize the behavior of the traffic flow in a complex network. Then we introduce our traffic model, which is based on the simple communication models on complex networks used in the literature. We perform a group of experiments to analyze the changes on the value of the scaling exponent when we modify the sampling process and the dynamical parameters of the model. Finally, we prove that many real networks do not fit in the two universal classes by analyzing the traffic of the Internet 2 backbone.

1. Scaling of fluctuations on complex networks

In a couple of recent articles, Menezes and Barabasi proposed a model to understand the origin of fluctuations in traffic processes in a number of real world systems, including the Internet, the world wide web, and highway networks (de Menezes and Barabási, 2004a, de Menezes and Barabási, 2004b). All of these systems can be represented at an abstract level as networks in which packets travel from one node to another, packets being real data packets or bits in the Internet, files in the world wide web, and vehicles in road networks. With the available resources nowadays, the movements of this packets can be measured simultaneously in all the nodes, obtaining a multiple time series description of the traffic flow (as presented in figure 4.1).

Due to the large amount of data available and the complexity associated to the dynamical process, an statistical analysis of this time series seems a good choice to obtain a characterization of the global behavior of the system. In particular, Menezes and Barabasi considered the relationship between the average number of packets $\langle f_i \rangle$ processed by nodes during a certain time interval, and the standard deviation σ_i of this quantity. Plotting the value of the dispersion as a function of the average traffic for all the nodes, they observed a power law scaling relationship,

$$\sigma \sim \langle f \rangle^{\alpha} \tag{4.1}$$

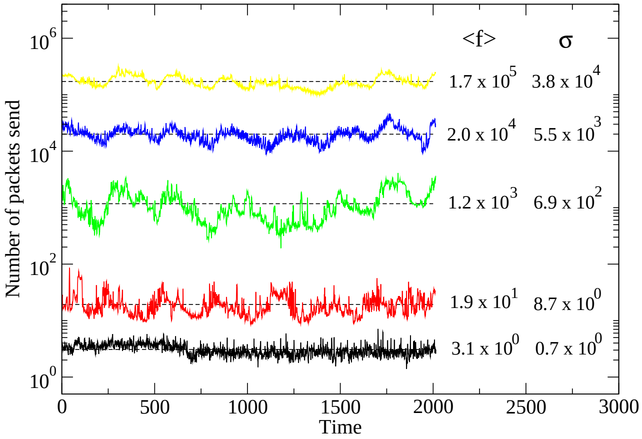


Figure 4.1. Left: Example of the traffic flow that goes through five routers that belong to the Internet2 network gathered from the Abilene project (see text for more details). On the right of each time series we show its average flow and the dispersion.

where α refers to the scaling exponent. They find that there are two classes of universality in this relationship for real systems. In the Internet, σ scales as $\langle f \rangle^{1/2}$, whereas σ scales as $\langle f \rangle$ for the world wide web and highway networks. Based on a stylized model of random walkers throughout the network, they conclude that this difference is due to the fact that the dynamics of the Internet is dominated by "internal noise" whereas the dynamics of the world wide web and highway networks is dominated by the demands of users, that is "external noise".

One of the main critics to their work refers to the simplicity of the model used to prove their theories. In the abstraction process proposed by the authors, they overlook what is probably one of the most important factors in the dynamics of traffic on networks, the limited capacity of nodes to handle packets simultaneously, which results in packet-packet interactions and, eventually, in large fluctuations or even network congestion (Guimerà et al., 2002b, Tadic et al., 2004).

2. A simple traffic model

To understand better the origin of the scaling relations for the fluctuations in networks, let us consider the behavior of a single node (for example, a toll plaza in a highway) trying to satisfy demands from users (vehicles arriving to the toll). As we learn from queueing theory (Allen, 1990), two stochastic processes fully determine the behavior of the node: (i) the arrival process by which new packets arrive to the node, and (ii) the service process by which the node satisfies the demands of the users, that is, forwards the packets. The most common queue model corresponds to the M/M/1 queueing system, where the randomness of the packets generation assumes a random (Poisson) arrival pattern and the service distribution assumes a random (exponential) time. The communication process in the case of a M/M/1 queuing system for each node in a complex network is well described by the so-called Jackson networks (Jackson, 1957).

Taking into account these considerations, we propose to model the traffic process in a complex network of N nodes as N queue systems of type M/M/1, and a random walk simulation for the movement of packets on the network. The arrival process of packets to the network is controlled by a Poisson distribution with parameter ρ, and each packet enters the network at a random selected node. Once the packet arrives to the node enters a queue. The delivery of the packets in the queue is controlled by an exponential distribution of service times with parameter μ. In our model, the packets will perform S random steps in the network before disappearing, being then S a measure of the persistence of packets in the network. This dynamics is performed in continuous time, assuming that the time expended by packets traveling through a link is negligible.

The system achieves a stationary state whenever the arrival rate of packets at each node is lower than or equal to the delivery rate, otherwise the system congests. The arrival rate at each node i is dependent on the topology and follows a distribution whose mean is $\rho_i^{ef} = \mathcal{B}_i\rho$ where $\mathcal{B}_i$ is the algorithmic betweenness of node i. $\mathcal{B}_i$ is defined as the relative number of paths in the network that go through node i given a specific routing algorithm (Guimerà et al., 2002b). As a direct consequence, the node with maximum algorithmic betweenness $\mathcal{B}_*$ determines the onset of congestion.

This traffic model is unable to reproduce the self-similarity of traffic in time observed in some real systems, as for example the Internet. It has been discussed that Poisson models aren't realistic (Leland et al., 1995) because do not reproduce some characteristics of the real dynamics like 'burstinness' that Internet exhibit. However, there are some authors (Karagiannis et al., 2004) that still defend that in certain cases Internet traffic can still be modeled using Poisson models, mainly when we are near the edge of congestion.

3. Effect of the dynamical parameters on the scaling exponent

In the following experiments we will focus on the average number of packets $\langle f_i \rangle$ processed by nodes during a certain time window of length P, and the standard deviation σ_i. The simulation of the dynamical process has been performed in a scale-free network with exponent for the degree distribution $\gamma = 3$ of 1000 nodes. We have observed the same results for larger SF networks,

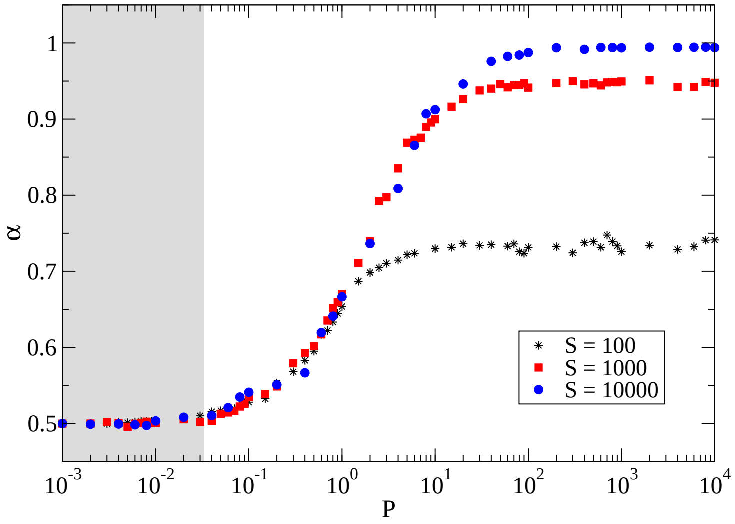


Figure 4.2. Value of the exponent α versus the time window length P in which averages are performed, for a fixed $\rho_*^{ef} = 1/3$ and different values of the persistence of packets in the network S. The shadowed area highlights the region of P in which the exponent $\alpha = 1/2$ always appears.

however the computational cost for the whole set of parameters used in the experiments becomes prohibitive.

3.1 Effect of the time window

The first parameter that we have studied is the effect of the size of the sampling window. Selecting a value of $P \ll 1/\rho_*^{ef} = 1/(\mathcal{B}_*\rho)$, we will always observe the scaling $\sigma \sim \langle f \rangle^{1/2}$, regardless of other parameters.

The explanation for this phenomena is very intuitive: Due to the value of P selected, the nodes will deliver either one packet or none, at each time interval. Suppose that during a number n_1 of intervals of length P the node deliver a packet whereas it does not deliver during a number of intervals $n_0 = n - n_1$, where n is the number of samples for the statistics. In this situation we also have $n_0 \gg n_1$. Therefore, the average and the standard deviation read

$$\langle f \rangle = n_1/n \tag{4.2}$$

$$\sigma = [\frac{1}{n}\left[n_1(1 - \langle f \rangle)^2 + n_0\langle f \rangle^2\right]]^{1/2}$$

which can be simplified to

$$\sigma = [(1 - \langle f \rangle)\langle f \rangle]^{1/2} \tag{4.3}$$

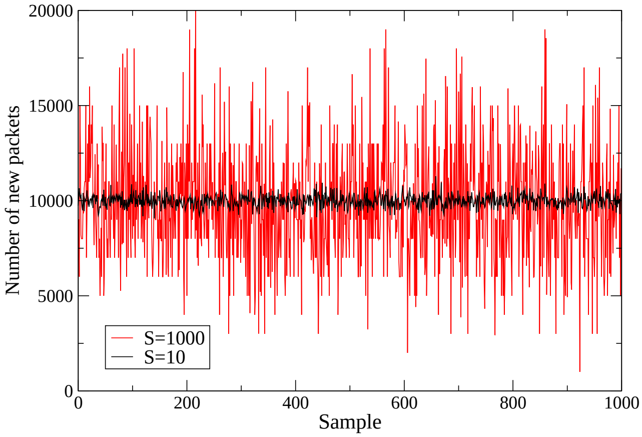


Figure 4.3. Example of two packet injection rates with the same mean but different variability. This has been done by varying the values of ρ and S proportionally

But, in the current scenario, the average flow is $\langle f \rangle \ll 1$ and then we recover the $\sigma \sim \langle f \rangle^{1/2}$ scaling law. Otherwise, this argument cannot be applied, and the scaling value will be influenced by the rest of parameters of the model.

In figure 4.2 we show the behavior of the scaling exponent α as a function of the time window length P in which the averages were taken, for a fixed $\rho_*^{ef} = 1/3$. We observe (shadowed area) that the exponent is always $1/2$ when the interval length is small enough. Indeed, from the data used the exponent $1/2$ stands for values of P.

The effects of the time window have been revisited in (Eisler et al., 2005). In this case, they observe that for larger time windows there is another transition between exponents, which is provoked by the existence of autocorrelations in the time series that only appear when the time window is large enough.

3.2 Effect of the traffic variability

Let us now assume that the sampling of the data is performed at intervals of length $P \gg 1/\rho_*^{ef}$. In this case, we expect the scaling of fluctuations in the system, beyond the effect of the sampling process, to be revealed. We analyze the behavior of the system varying the rate of injection of packets into the system ρ and the number of steps S each packet performs before it disappears. We first consider that the service rate $\mu \to \infty$. In this case, the effect of queues is minimized and then no interaction between packets is accounted for. The total traffic $\mathcal{T}$, number of packets flowing through the network per unit time, is determined by the Poisson process with mean $\langle \mathcal{T} \rangle = \rho S$. Keeping the total traffic mean $\langle \mathcal{T} \rangle$ fixed, we can control the variability of the local traffic incoming to a node by varying the values of ρ and S proportionally. Figure 4.3

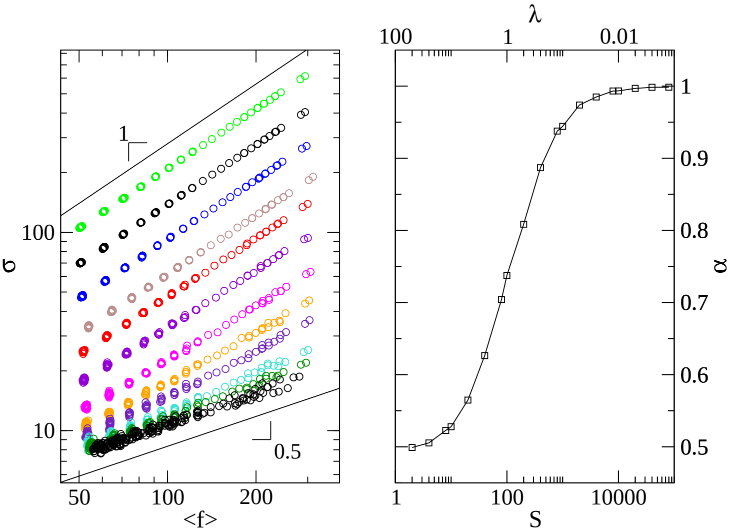


Figure 4.4. Left: Plot σ versus $\langle f \rangle$ for different realization of ρ and S maintaining its product constant. Right: Plot of the α exponent for $\rho S = 100$. Other values of ρS have produced equivalent results, shifted to a different region of $\langle f \rangle$.

shows the differences between a traffic with low variability (small S) and large variability (large values of S) with the same mean average.

In figure 4.4 we show the scaling exponent transition between $\alpha = 1/2$ and $\alpha = 1$. This plot recovers the results depicted in (de Menezes and Barabási, 2004a), although the explanation should be reconsidered in the new scenario. The transition of exponent from $\alpha = 1/2$ to $\alpha = 1$ is obtained here simply by increasing the number of steps S the packet performs on the network while maintaining the mean value of the total traffic (i.e. decreasing proportionally the injection ratio ρ).

This results contradict the interpretation in (de Menezes and Barabási, 2004a) because increasing the number of steps in the network increases the internal fluctuations of traffic because more packet-packet interaction occurs, while decreasing the injection of packets (remember, Poisson distributed) decrements the external fluctuations of traffic in this scenario. Nevertheless both results are coherent at this point concerning the scaling of fluctuations. Our interpretation of this transition is the following: for the same total traffic on the network, the nature of fluctuations is related to the number of steps S the packets perform on the network. When the number of steps is small enough the behavior of fluctuations is akin a random deposition process independent of the topology of the network, $\rho_i^{ef} \approx \rho$. When the number of steps in the network grows, the

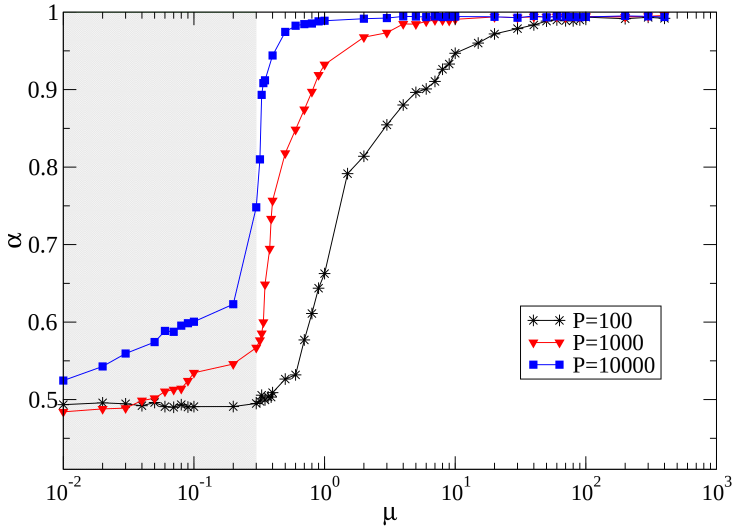


Figure 4.5. Scaling exponent α as a function of the time service μ, for three different time window lengths, and for $\rho_*^{ef} = 1/3$. Shadowed area highlights the region where congestion starts at nodes with $\rho_*^{ef} = 1/3$.

topology induces dynamical correlations that affect the scaling of fluctuations via the algorithmic betweenness, $\rho_i^{ef} \approx \rho\mathcal{B}_i$.

3.3 Effect of the congestion

We extend the simple model where queues are neglected, to the more realistic situation when queues are persistent. The introduction of queues in the system, in our model, is controlled by the parameter μ (rate of service). The possible values of μ are constrained by the onset of congestion i.e. $\mu > \rho_*^{ef}$, otherwise congestion appears at those nodes with $\mathcal{B}_*$, because of the arrival of more packets than those that can be delivered. We investigate those values of μ near the onset of congestion to reveal the effect of queues in the scaling properties of the system.

When congestion occurs, the queues corresponding to those nodes with $\mathcal{B}_*$ will have always more packets that those than can be delivered in a period P. That means that the number of packets delivered by these nodes will be controlled exclusively by the service rate μ, i.e. the variance scaling with respect to the mean flow at these nodes will be again fitted by $\alpha = 1/2$ corresponding to the exponential service distribution. Close to the onset of congestion we approach the situation where the scaling exponent $\alpha = 1/2$ should be recovered, however the possibility that in some periods of time the queues will be unoccupied increases as we go away from the congested regime, thus a new transition in the scaling exponent as a function μ is expected. In figure 4.5 we plot the scaling exponent transition as a function of μ for a fixed value of $\rho_*^{ef} = 1/3$. In this situation the onset of congestion is determined by the critical value $\mu_c = 1/3$. Note that for values below μ_c some nodes of the network collapse and then gradually the rest of the nodes in the network. In this region, shadowed area of figure 4.5 the system enters the congestion regime progressively. The transition on the scaling exponent depicted in figure 4.5 is also affected by the time window length P, we plotted the transition for $P = 10^2$, 10^3 and 10^4. We observe that as P increases, the transition becomes sharper. Indeed in the limit of $P \to \infty$ we conjecture that the transition could be discontinuous, and could reflect a first order phase transition as observed in other traffic models (Echenique et al., 2004), although we can not claim that this discontinuity will occur sharply from 1 to $1/2$.

4. Scaling exponent of Internet traffic

Up to now, we have show that a simple traffic model where the injection of packets to the system follows a Poisson distribution, can account for different scaling exponents α depending on the parameters ρ, μ, S and the time period P were the statistics are performed. These results lead us to suspect that the scaling of fluctuations in real systems must be affected by these parameters as well. This cast doubts on the universality predicted in (de Menezes and Barabási, 2004a). Indeed, this non-universality has been also claimed in the exponent of fluctuations when studying the data flow between stocks in NYSE market (Eisler et al., 2005), or in the e-mail activity of one user (Eisler et al., 2007).

To corroborate our doubts about universality on the scaling of fluctuations in complex networks, we have studied the Internet traffic between routers of the Abilene backbone network that are part of the data also used in (de Menezes and Barabási, 2004a). The Abilene network is the U.S. high-performance backbone network created by the Internet2 community as a testing environment in 1999 (see figure 4.6). Since then it has been publishing a large amount of information about its performance, including the amount of traffic that passes trough each router interface[1].

We collected data from the 112 available router interfaces (links). We gather information of the number of packets that exit through each router interface between September 15th and November 15th of 2005, at intervals of 5 minutes. The scaling $\sigma \sim \langle f \rangle^\alpha$ shows exponents that range from $\alpha = 0.71$ to $\alpha = 0.86$, significantly different from the exponent $1/2$ presented in (de Menezes and Barabási, 2004a).

[1] This information is publicly available at http://abilene.internet2.edu.

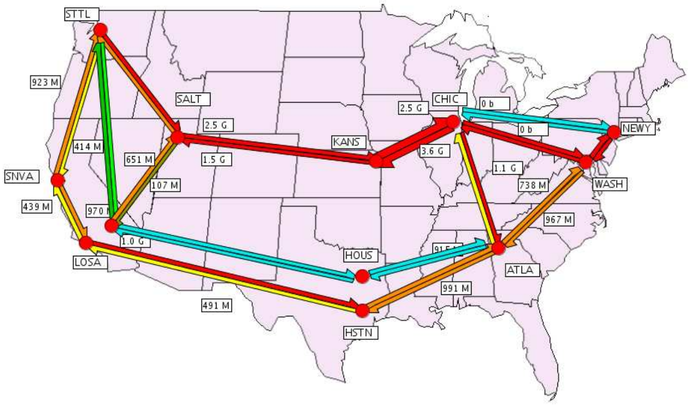


Figure 4.6. Map of the topological structure of the Abilene network. This network acts as the backbone infrastructure of Internet 2, connecting a large number of administrative, educative and private corporations. Map downloaded from http://abilene.internet2.edu/maps-lists/.

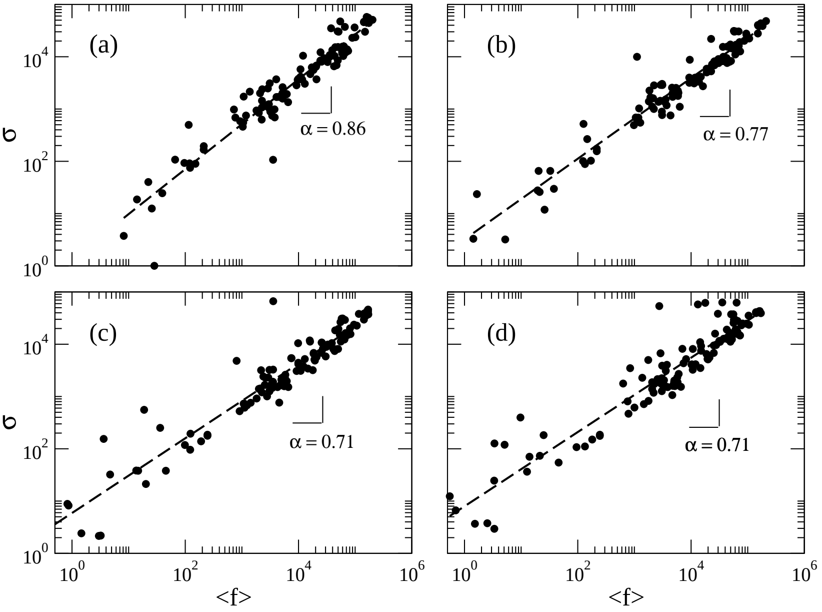


Figure 4.7. Scaling relations between σ and $\langle f \rangle$ for the 112 Abilene backbone router interfaces. Analysis performed during (a) two days, (b) one week, (c) one month and (d) two months, finishing all them in November 15th of 2005. The time window length P is fixed to 5 minutes.

The interpretation of these exponents in the context of our stylized model is that the Abilene backbone is far from the onset of congestion for the interface with maximum algorithmic betweenness, and seems compatible with the mean rate of utilization of the interfaces in this backbone that is usually below 30%.

5. Summary

In this chapter we have presented a simple model of traffic in complex networks that capture the essential parameters governing the dynamical process. The model shows a scaling relationship between σ and $\langle f \rangle$ whose exponent depends on the parameters considered as well as on the time window in which the statistics are performed. Moreover we have shown that the corresponding exponent for the scaling of fluctuations in the Internet Abilene backbone network is different from $1/2$ as stated in previous works, corroborating by exclusion that the universality on the scaling of fluctuations in complex networks should be questioned.

The next logical question should be then, if there is not universality to explain the origin of the fluctuations, what determines the exponent for each real system? Here we have presented that many factors can control the behavior of the fluctuations, but still we need to determine the specific reasons of each exponent in real networks. Moreover, we will probably determine other causes that can provoke more other transitions between $1/2$ and 1.

This work opened the door to another group of studies that have focused on the influence of the topology on the fluctuations. There is still a large number of experiments and theories that we can perform about the study of fluctuations. A possible extension of this work will be the use of a generation model that reproduce the self-similarity expected, and then study the exponents obtained by this new traffic model. We guess that the self-similarity will be reproduced if the injection of packets into the system follows a heavy-tailed distribution instead of a Poisson distribution, however we can still not prove this conjecture. This will also open the door to the study of the relationship between the values of the exponent with the self-similarity using the Hurst exponent.

Chapter 5

DYNAMICAL ROBUSTNESS OF A COMMUNICATION PROCESS

In this final chapter we will focus our attention in another interesting property of many complex networks, its resilience (or robustness) to the failure of some of their nodes. The robustness plays a key role maintaining the functionality of the dynamic processes that take place in a complex network. In the case of the Internet, the stability against node failures is a key factor to maintain the performance and the efficiency of the network (which is reflected in low packet loss rates and short packet traveling times).

Traditional studies have analyzed the effects of topological percolation in complex networks, proving differences between classes of complex networks when they undergo attacks or random failures. Most of the studies define the robustness of a network as its capability of maintaining most of its nodes connected, forming a giant component of the same size as the original network. But in real complex networks, an interesting process happens before a connected network splits, namely that even though the underlying network is still connected, the dynamical processes taking place on it can change significantly due, for example, to congestion effects. In this scenario we will introduce the concept of dynamical robustness of a network, defined as their capacity to continue working when some of the nodes fail.

The purpose of this chapter is the analysis of the dynamical robustness of a communication process. Using a similar traffic model to the presented on the previous chapter, we will measure the effect of a random node removal on the onset of congestion. The chapter is organized as follows: First we introduce the differences between the dynamical and the topological robustness of a communication process. Then we perform some experiments to determine the dynamical robustness by analyzing the changes on the maximum capacity of the network. To perform this task, we have simulated random failures in three different types of networks (regular, Erdös-Rényi and scale-free), and using a range of protocols with different radius of knowledge (from shortest paths to random walks). Finally we focus on the relationship between topological robustness and dynamical robustness by comparing if the network will be first physically split or dynamically collapsed.

1. Robustness of Complex networks

Many real complex networks display a high robustness against random failures (Albert et al., 2000). This phenomenon has been successfully related to their scale-free degree distribution (Cohen et al., 2002, Gallos et al., 2005); with a very high probability the random failures will affect the lowest connected nodes, which have small influence in maintaining its structural properties (Cohen et al., 2000). However, the same degree distribution is also responsible of the vulnerability of scale-free networks against directed attacks (removal of the most connected nodes) (Cohen et al., 2001).

Internet has been considered as a paradigmatic example of this “robust yet fragile” structure (Doyle et al., 2005). On one hand, everyday a large amount of nodes suffer temporal failures without affecting the global behavior, since the overall system is able to redistribute the traffic while there is a path connecting the elements. On the other hand, this robustness coexists with a fragility of its central elements to fail under a malicious activities. A directed attack against a specific selected nodes can decrease the efficiency of the network, even disconnect it in two or more components[1].

Several studies have covered the incidence of a node removal on the statistical properties of complex networks, such as the diameter (Albert et al., 2000), the average path length (Holme et al., 2002b, Gallos et al., 2005) or the size of the giant component (Albert et al., 2000, Callaway et al., 2000). Since these properties play an important role in the interplay between the topology and dynamics of complex networks, the node removal will also change the dynamical processes supported on the network (Tadic et al., 2007).

One of the properties that is affected by the removal of nodes is the *efficiency* of nodes to distribute traffic in communication processes (Latora and Marchiori, 2001, Crucitti et al., 2003, Lopez et al., 2007). The efficiency between nodes i and j is defined as the inverse of the shortest path connecting them. This property, related to the information flow in networks, is interesting because it allows to quantitatively compare the dynamical performance in traffic of different network structures, however, it obviates one of the most important aspects of any communication process: congestion. In real networks, each node has a limited capability to deliver information, meaning that they can serve a bounded number of “packets” of information per unit time. When

[1] For instance, some failures of the transatlantic communication routers have temporary split the connection between Europe and America, breaking the giant component of Internet into two or more parts.

the incoming traffic exceeds this capability, the system enters a congested state, there is no balance between incoming and outcoming traffic, and the communication processes become inefficient (Guimerà et al., 2002b).

A typical example of the effects of such a breakdown is found in power grid networks. The removal of a certain fraction of nodes triggers a cascade failure on the system (Motter, 2004, Crucitti et al., 2004). This failure is caused by the redistribution of the traffic flow between the remaining nodes, surpassing their capability and therefore collapsing some of them. This cascade phenomenon has also been observed in the Internet, where the failure of one router can trigger additional failures due to the redistribution of the traffic, which may generate a congestion collapse which will avoid the connection between a large group of elements (Holme and Kim, 2002, Moreno et al., 2003).

2. Determining the Robustness of a Communication Process

At the light of this results, seems clear that the communication between two nodes can be interrupted by two different causes: if the network physically splits or if the traffic can not be delivered due to the existence of congestion. To model this two causes, from now on we will differentiate between the topological and the dynamical robustness of a network. The first is related to the process of node removal and its topological effect. The second is related to the changes on the onset of congestion for the traffic dynamics when the removal of nodes is performed.

2.1 Topological robustness

A random breakdown of a network can be modeled as a percolation process. The percolation threshold p_c in lattices is defined as the fraction of lattice points that must be filled to create a continuous path of nearest neighbors from one side to another, or equivalently destroyed to ensure that no such a path exists. In complex networks, the percolation threshold is usually characterized by the existence of a giant component with the same diameter as the original network (Albert et al., 2000, Holme et al., 2002b). The diameter keeps constant while the size of the giant component is $\sim O(S)$, being S the original size of the network.

In this chapter we use a more restrictive approach to determine the percolation threshold akin to that used in lattices. Instead of considering the size of the giant component, we will look for the critical fraction of node removals that avoids the existence of a physical path connecting every pair of the remaining nodes of the network.

The topological robustness of a network is defined then as the probability of maintaining all nodes connected when increasing the fraction p of removed

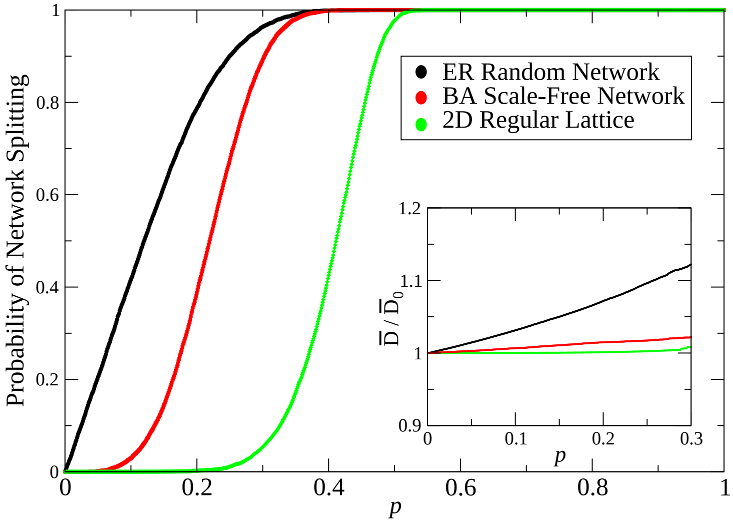


Figure 5.1. Probability of network splitting in two or more connected components as a function of the fraction of removed nodes p. Inset, relative average path length as a function of p.

nodes. For $p < p_c$, the probability of having more than one component is zero. For $p > p_c$ the probability shows a transition determined by the statistical properties of the network. We have studied this robustness on three different types of networks: regular lattices, ER and scale-free. For comparison purposes, the three types of networks will have the same number of nodes $S = 1000$ and a similar number of links $L \sim 4000$.

The first type of networks have been implemented as periodical two-dimensional regular lattices with all nodes having the same degree $k = 8$. This type of networks have a very high clustering coefficient and a high mean average path length. The random networks have been created using the ER model with an edge probability $p = 0.008$. In this case, the networks display a Poisson degree distribution with a mean value of $\langle k \rangle = 8$, a very low clustering coefficient and also a low average path length. Finally, the scale-free networks have been created using the preferential attachment mechanism of Barabási-Albert (BA) where each node adds $m = 4$ new links, obtaining power-law degree distribution $P(k) \sim k^\gamma$ with $\gamma \sim 3$.

To calculate the topological robustness defined above, we have performed a sequential random degradation process, removing sequentially nodes (and their connections) until the network splits. We have repeated the breakdown 10^6 times to obtain a significant statistical approach.

Figure 5.1 shows the probability of splitting the network (topological robustness) when a fraction of nodes p have been removed. The results are similar to the fragmentation processes exposed in previous articles (Albert et al., 2000, Cohen et al., 2000) . The probability threshold is lower when the network has a power-law distribution due to the existence of hubs that act as cohesive elements, hardly destroyed by a random process. As a consequence, the network remains connected for larger values of p compared to ER networks. The reason for the robustness of regular networks is different, the high robustness exhibited is a consequence of their high clustering coefficient which provides a high degree of redundancy. In the inset we plot the relative average path length as a function of the number of removed nodes, as expected the average path lengths remain almost constant in all networks, with a slight increment in the ER networks case. These results in ER and BA networks are in agreement with results of random percolation in complex networks (Albert et al., 2000, Cohen et al., 2000, Cohen et al., 2002), showing however a shift in the transition point due to a more restrictive definition of robustness used here.

2.2 Dynamical robustness

In analogy of the topological robustness, the dynamical robustness is defined as the probability of a network to maintain the communication processes between every pair of nodes nodes. When the network splits into two or more components, the communication is also interrupted because the packets are unable to reach the isolated nodes. However, there are some cases where the network is still physically connected but the overlay dynamics is unable to deliver information to certain nodes, provoking some sort of dynamical split of the network. As we have introduced before, the cause of this phenomenon is the emergence of congestion. When a system is congested, a large number of packets get stuck in nodes and, if there are delivering time restrictions, never reach their destination.

To study the congestion point we will use a similar traffic model to the presented in the previous chapter: First, to model the receiving and limited transmission of information of each node, we assign a queue to each one and a different from zero service time. The capability of the nodes of is characterized then by the time needed to serve one packet. We assume this time to follow an exponential distribution with mean $1/\mu$. If a packet arrives when the node is busy delivering another one, it gets stored in a FIFO (first in-first out) queue until it gets dispatched. To simplify the experiments, we will use during the rest of the work a value of $\mu = 1$.

Once we have mapped the queues into the nodes, we introduce the dynamical rules: The packets are created in each node following a Poisson distribution with mean ρ, and they are assigned with a random destination. These packets travel through the network using a static routing protocol (the decision rules are set at the beginning of the experiment). Once the packet arrives at its destination, it is removed from the system.

Since congestion emerges when the incoming traffic to a node is higher than its capability to dispatch it, and we have fixed the value of this capability by the service time, the onset of congestion remains as a function of ρ. The network achieves its steady state when for a certain value of ρ the number of packets of the system at time t, $N(t)$, fluctuates around an stationary value. When the value of ρ overcomes a critical value ρ_c, the number of packets $N(t)$ diverges and the system enters in a congestion phase. Moreover, it has been proved (Guimerà et al., 2002b) that the onset of congestion is driven by the node with the highest algorithmic betweenness B^*. The algorithmic betweenness of a node B_i is the number of paths that go through node i given a certain routing algorithm. When the incoming traffic that arrives to this node is higher than its delivery capability, $\rho B^*/(S-1) > \mu$, its queue starts to grow and induces congestion in the network. Therefore, the congestion point of the system ρ_c is determined by the moment at which the node with maximum algorithmic betweenness receives and delivers the same ratio of packets:

$$\rho_c = \frac{\mu(S-1)}{B^*} \tag{5.1}$$

2.3 How to determine the dynamical robustness?

Our experiments to determine the dynamical robustness consist then in to analyze the variation of the onset of congestion determined by ρ_c when the system experiments random failures, simulated as the sequential random elimination of nodes, for those networks that after the random failure still remain connected. For each network, we perform a step of the sequential removal of nodes, if the removal of the node does not produces a split on the network we calculate its new ρ_c.

To determine numerically the value of ρ_c for a given configuration we simulate the traffic dynamics. Starting from a value of ρ that provides a steady state, we gradually increase this value and determine whether or not the number of packets floating on the system diverges. The difficulty of deciding if the system is or not at the critical point, increases as ρ approaches ρ_c. To characterize the transition we used an order parameter η (Arenas et al., 2001):

$$\eta = \frac{N(t+\tau) - N(t)}{\rho\tau} \tag{5.2}$$

where τ is the observation time. When $\rho < \rho_c$ the order parameter is zero (There is no difference between the ratio of created packets and the ratio of removed). On the contrary, if $\rho > \rho_c$, the value of $N(t)$ grows linearly with t, and the order parameter is a function of ρ.

Before removing nodes, we determine the maximum load that the complete network can handle $\rho_c(0)$. Then we remove a fraction of nodes p and recal-

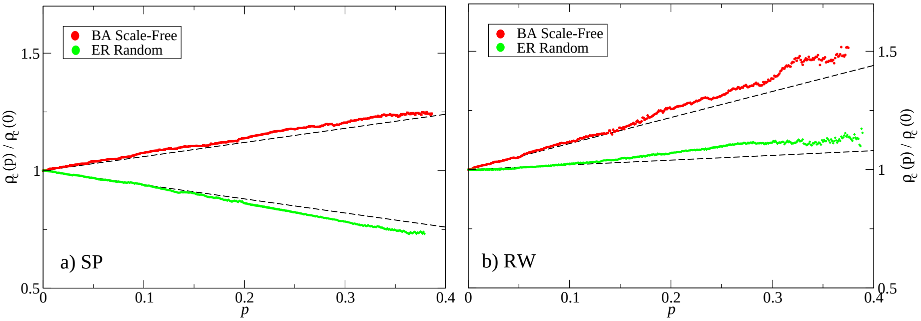


Figure 5.2. Effects of the node removal on the normalized maximum capacity of scale-free and random networks when using a (a) routing protocol based in shortest paths and (b) routing protocol based in random walks. Dotted lines present the analytical approach using equation 5.5 and experimental data from table 5.2.

culate the maximum congestion value $\rho_c(p)$, repeating this process while the network has more than one connected component. We perform 10^4 simulations of each experiment to obtain an statistical approach of $\rho_c(p)$.

3. Effects of a node removal on the onset of congestion

We have performed three different experiments, trying to understand the changes on the congestion point for different topologies and routing protocols.

3.1 ER and SF networks

In a first experiment we have analyzed the effects of the random breakdown on congestion, when the movement of the packets is governed by a shortest path (SP) routing protocol. The results obtained are shown in figure 5.2 (a). We observe different behaviors depending on the topology used: The maximum load in a scale-free network increases with the number of removed nodes, whereas in ER random networks decreases slowly with p.

To understand this results we have studied the changes on the betweenness distribution of both network structures when a node is removed. It has been proved that there is a correlation between the degree and the betweenness distribution in random graphs and in scale-free networks (Holme et al., 2002a, Goh et al., 2001). Since the probability of deleting the node with the highest degree is very small, we can consider as a first approximation that the node with B^* is the same during all the breakdown process.

Another important feature that we can extract from the betweenness distribution is the importance of one node in the communication process (Barthelemy, 2004, Latora and Marchiori, 2004). We characterized the importance of the most important node using α^*, defined as the maximum algorithmic betweenness normalized, $\alpha^* = B^*/\sum B_i$. We can see in table. 5.2 that the importance of the most connected node α^* differs significantly in the random and scale-free networks due to their different degree distributions: In the scale-free, 7% of the packets travel through the most central node in contrast with the the $0,52\%$ of the random network.

Every time we remove a node i, we also remove the load it generates L_i (Motter, 2004), decreasing the value of the B^* according to the importance of this node in the communication process. The load generated by one node is defined as

$$L_i = \sum_i (D_{i,j} + 1) = (\bar{D}_i + 1)(S - 1) \tag{5.3}$$

where $\bar{D}_i$ is the average path length between node i and the rest of nodes in the network $(S - 1)$ (being S the original number of nodes of the network). In a SP routing protocol, this distance measures the average shortest path length from node i to the rest of the network, which can be easily determined using a Dijkstra algorithm (Cormen et al., 1990) . Using equation (5.1) we express the onset of congestion for a certain fraction of removed nodes $\rho_c(p)$ for large S as:

$$\rho_c(p) = \frac{(S-1) - pS}{B^*_{ini} - \alpha^* \bar{L}(p)} \sim \frac{S(1-p)}{B^*_{ini}(1 - \frac{\alpha^* \bar{L}(p)}{B^*_{ini}})} \tag{5.4}$$

where pS is the number of removed nodes and $\bar{L}(p)$ is the amount of load that we have withdrawn of the network after deleting pS nodes, which can be approximated by $\bar{L}(p) \sim pS\bar{L}$, where $\bar{L} = \frac{1}{N}\sum_i L_i$. equation 5.4 can be approximated using a Taylor expansion, obtaining

Network	Protocol	B^*_{ini}	α^*	$\bar{D}$	$\bar{L}$	$S\bar{L}\alpha^*/B^* - 1$
BA Scale-free	Shortest Path	$1.5 * 10^5$	0.07	3.3	3300	0.54
ER Random	Shortest Path	$1.3 * 10^4$	0.002	3.8	3800	-0.52
BA Scale-free	Random Walk	$2.2 * 10^7$	0.029	1595	$1.6 * 10^6$	1.1
ER Random	Random Walk	$2.9 * 10^6$	0.0024	1380	$1.4 * 10^6$	0.15

Table 5.1. Values of the maximum algorithmic betweenness B^*, the importance of this betweenness in the communication process α^*, the average path length $\bar{D}$, and the average generated load by one node $\bar{L}$ for the scale-free and random networks using SP and RW routing protocols. The value of $S\bar{L}\alpha^*/B^* - 1$ determines the change of the congestion point when removing a fraction of nodes p.

$$\rho_c(p) \sim \frac{S}{B^*_{ini}} + p\frac{S}{B^*_{ini}}\left(\frac{S\bar{L}\alpha^*}{B^*_{ini}} - 1\right) + O\left[\left(\frac{pS\bar{L}\alpha^*}{B^*_{ini}}\right)^2\right] \tag{5.5}$$

Considering that $pS\bar{L}\alpha^*/B^*_{ini} << 1$ when $p << 1$, using equation 5.5 we can determine the expected initial behavior of the congestion point analytically. When $S\bar{L}\alpha^*/B^* > 1$ the maximum load supported by the system starts to grow as the node suffers a random removal, and the initial slope of the congestion is $S\bar{L}\alpha^*/B^* - 1$. Otherwise, if $S\bar{L}\alpha^*/B^* < 1$ the maximum load decreases with the node removal. Introducing the values presented in table 5.2 in equation 5.5, we have represented in figure 5.2 the expected behavior of $\rho_c(p)/\rho_c^{ini}$, obtaining a good agreement with the computational simulations.

We have also analyzed the ratio $\rho_c(p)/\rho_c(0)$ when packets are delivered using a random walk (RW) routing protocol (Noh and Rieger, 2004). The RW betweenness distribution for a random walk process has been studied in (Newman, 2003a), showing that it shares the properties of the SP betweenness. table 1 shows that when we use a RW routing protocol the statistical values increase significantly. The average path length of a packet to reach its destination is much higher than the shortest path, since the packets do not have information about how to reach their destination. This distance can be determined analytically using the mean first-passage time between two nodes (Noh and Rieger, 2004). Since the distance is much higher, the amount of load introduced by the nodes is also higher, and therefore the value of B^* increases. The results obtained (see figure 5.2(b)) show that using a RW routing protocol, the initial congestion behavior is also governed by the evolution of the B^* described in equation 5.5, although a larger deviation is observed for larger values of p in agreement with the discussion above.

3.2 Regular lattice

We have performed a second experiment to investigate the behavior of the congestion when the underlying topology is a regular lattice, see figure 5.3. This type of network is interesting because the changes on the congestion can not be described in the previous approximation.

The explanation for the behavior observed in figure 5.3 is the following. Before removing any node of the regular network, all of them have the same algorithmic betweenness because the underlying symmetry. When a little fraction of nodes has been removed, the shape of the betweenness distribution changes, and some nodes become more relevant in the communication process. The changes of this centrality are characterized by the changes of α^*. In the precedent analysis we have considered that the value of α^* is constant because the failures does not modify significantly the structural properties. However, in the regular network this process change the structure breaking symmetry, and

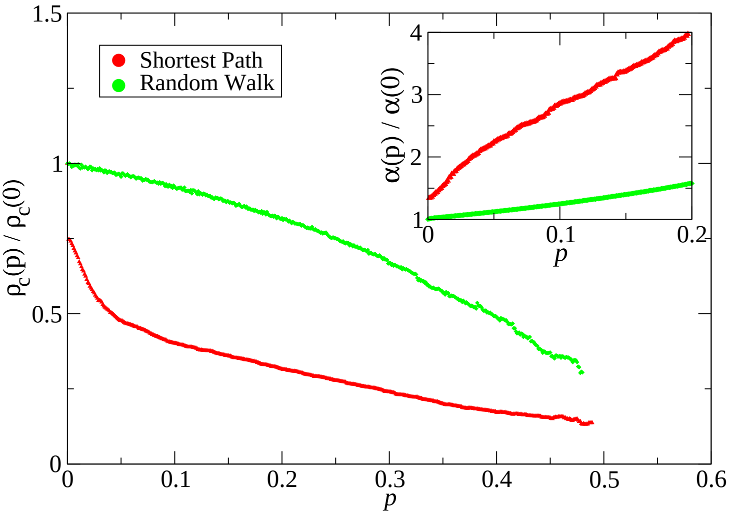


Figure 5.3. Effects of the node removal on the normalized maximum capacity of a 2D regular lattice networks when using a routing protocol based in shortest paths and random walks. In the inset we plot the evolution of the relative importance of the node with maximum betweenness as a function of p.

the value of α^* becomes now a function of the number of removed nodes. The inset of figure 5.3 shows the evolution of this parameter when we remove a certain fraction of nodes. Since the value of the maximum betweenness is a function of α^*, when this value grows B^* also grows and the onset of congestion decreases.

3.3 Effect of the knowledge radius

Up to now we have observed different type of behaviors of the onset of congestion when we perform a random breakdown. Particular interest raises the observation of the ER network (figure 5.2 right), where the behavior depends on the routing protocol: when shortest path are applied, the capability decreases, however when random paths are followed the capability increases. Therefore, seems interesting to analyze the transition between this two different protocols, introducing a certain amount of neighboring information in the routing decisions.

In the third experiment we have analyzed what happens when we use a routing protocol with different values of the knowledge radius r, representing a system whose nodes have a limited knowledge of network topology determined by the radius. The use of a knowledge radius is found in many real systems, where due to space limitations the elements only know the exact location of a few nodes, and otherwise they can only guess the position of the

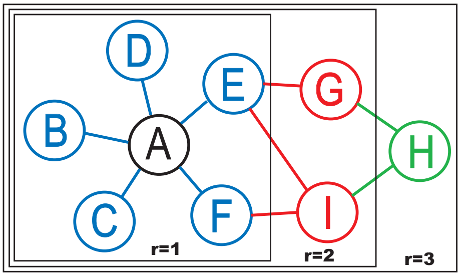


Figure 5.4. Example of a routing protocol with a certain knowledge radius r. When $r = 0$, node A does not know nothing about the location of other nodes and sends all the packets using a randomly chosen link. When $r = 1$, node A can send directly the packet to nodes in blue (B-F), and otherwise it send the packet randomly. When $r = 2$, know how to reach (B-I) and if $r = 3$ the node know the position of the rest of the nodes, and can send the packets using the shortest path.

destination node. In figure 5.4 we present an example of a routing protocol with a fixed knowledge of the network. When $r = 0$, the nodes do not know nothing about the topology, and the packets are always moved randomly. If we increase the value of r to 1, the nodes know who are its neighbors, and if the packet destination is one neighbor, the packet is sent directly to the destination. Otherwise the packet is sent randomly. For larger values of r, the probability of knowing the location increases, decreasing the number of random steps of the packets. An finally, when r is equal to the diameter of the network, all the node know exactly how to reach the destination, and all the packets travel using the shortest available path.

In this experiment we have repeated the same random removal process on the ER and the SF networks described previously, but now we use a knowledge radius that ranges between $r = 0$ and $r = 4$ (Which is approximately the average shortest path length of the networks). The results obtained are presented in figure 5.5. A first look at both plots shows an unexpected behavior of the onset of congestion as we increase the radius. For the value of $r = 0$ we observe the same behavior described in the previous section. But as we increase the value of knowledge radius, we observe that the slope of $\rho_c(p)/\rho_c(0)$ first decays to negative values (-1.14 for the ER network and -0.31 for the SF network when $r = 2$), and then increases again until it reaches the same slope for the shortest path presented in table 5.2.

To explain this unexpected results, we have studied how does the random removal changes the parameters that control the changes onset of congestion, instead of considering this parameters constant like in the previous section. Using the same methodology explained before for determining $\rho_c(p)$, we have analyzed the effect of the random the average number of steps that one packet performs before exiting the network $\langle D\rangle(p)$ and the maximum algorithmic

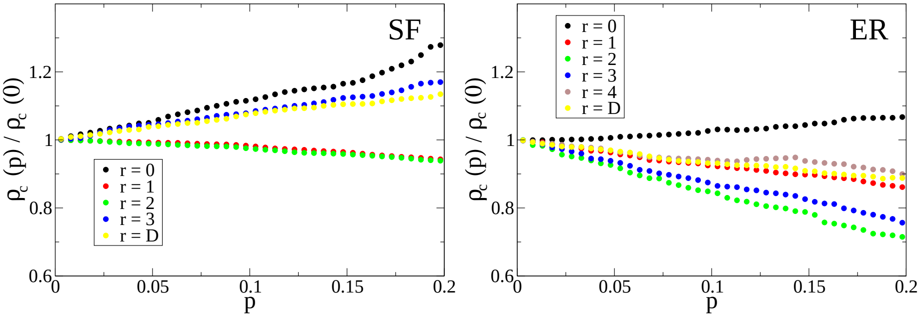


Figure 5.5. Effects of the node removal on the normalized maximum capacity of the ER (right) and SF (left) networks when using a routing protocol that depends on the knowledge radius of the network. The values of r range from 0 (Random Walk protocol) to the Diameter of the network (Shortest path protocol).

betweenness normalized $\alpha^*(p)$. Note that both values depend on the value of r.

A first look at the evolution of the average path length in figure 5.6 left gives the hint that it plays an important role in the changes on the value of $\rho_c(p)/\rho_c(0)$, since it repeats the same behavior of figure 5.5 but inverted. This is not surprising, since the fact of reducing (increasing) the distance means that the packets stay less (more) time in the network, and this increases (decreases) the maximum capacity of the network.

Let us analyze with detail the behavior of the changes of $\langle D \rangle(p)$ normalized. On one hand, we observe that in a purely random routing protocol the average path length decreases. In this particular case, the average number of steps is very high, $O(n)$, and the effect of removing nodes reduces the average number of nodes that we randomly explore before reaching our objective. On the other hand, when we introduce routing knowledge into the system, the average path length increases with different slopes. As explained in (Albert et al., 2000), this is because we remove a part of the nodes that can act as a shortest path, increasing the difficulty for the remaining nodes to communicate with each other. The effect is less pronounced in the SF network, since the most part of the paths go through the central hubs which usually are not removed.

We also have analyzed the changes on the maximum algorithmic betweenness of the network in figure 5.6 right, to check if the introduction of routing protocols reproduced the same effect observed in the regular lattice. In the scale-free network, the value of $\alpha^*(p)$ normalized remains almost constant, independently of the value of r, proving the stability of the most connected node to a random failure. In the random network there is a slight increase of the value that is similar for all the routing protocols, which can probably have some influence on the slope of $\rho_c(p)/\rho_c(0)$.

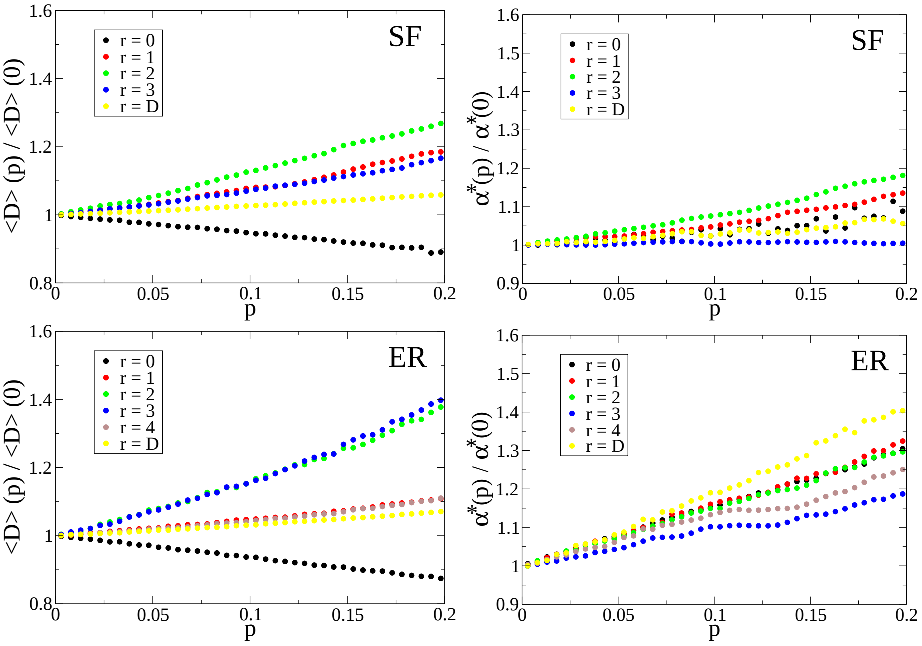


Figure 5.6. Effects of the node removal on some dynamical properties of a communication process as a function of the number of removed nodes p: the average number of steps that one packet performs before exiting the network $\langle D \rangle (p)$ and the maximum algorithmic betweenness normalized $\alpha^*(p)$. The top three figures refer to the Scale-Free network, and the bottom refer to the ER random network.

To confirm analytically all these observations we have computed a theoretical approach of the changes of the onset of congestion, using a modified version of equation 5.4,

$$\rho_c(p)/\rho_c(0) \sim \frac{(1-p)}{(1-\frac{\alpha^*(p)\bar{L}'(p)}{B^*_{ini}})} \tag{5.6}$$

where now the values of $\alpha^*(p)$ and the average distance used to compute $\bar{L}'(p)$ are the obtained in 5.6, instead of being constant like in the previous experiments. Figure 5.7 shows the comparison between the value of $\rho_c(p)/\rho_c(0)$ obtained in the simulations versus the theoretical approach obtained using equation 5.6. We observe a good agreement between the expected and the simulated values of $\rho_c(p)/\rho_c(0)$, confirming that the changes on the onset of congestion are governed by the changes on the distance and the centrality. Moreover, we can also observe that this approach acts as an upper bound for the changes of the onset of congestion.

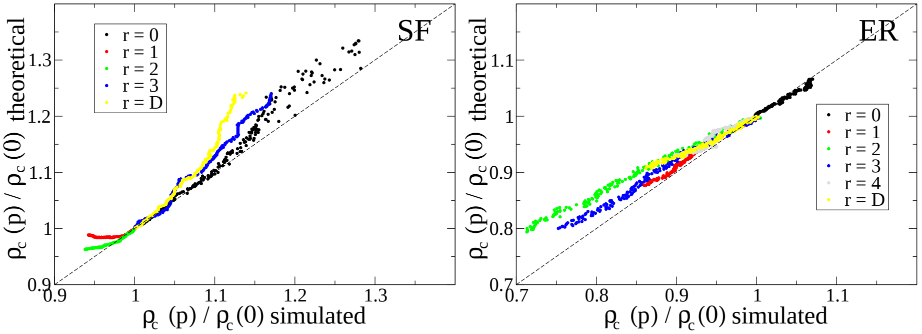


Figure 5.7. Comparison between the values of $\rho_c(p)/\rho_c(0)$ obtained using simulations and the values obtained using equation 5.4 and the data of figure 5.6. We observe that the numerical approach provides a good fitting of the behavior of the onset of congestion.

4. Comparing topological and dynamical robustness

As we have introduced so far, when a network suffers random failures there is the possibility that some nodes of the network get isolated from the communication process. The causes of this isolation can be topological, if the network splits, or dynamical, when congestion emerges and avoids the proper distribution of information. To discover which one of this two causes will appear first given a certain topology and routing protocol, we have compared the probability of disconnecting the network physically versus the maximum congestion of the network, when removing a fraction p of nodes. We use the three network topologies and the SP and RW routing protocols presented before, obtaining the results presented in figure 5.8.

This comparison provides some insights about the robustness of the communication process, defining regions of the parameters for which congestion is attained before splitting the network and vice versa. First we find that SF networks show a very high dynamically robustness. This means that, even if the system is functioning at his maximal capacity before removing any nodes, the random failures will not introduce congestion into this system. In second place, we find that regular networks are more topologically robust. If the communication process is based on a RW routing protocol and the initial system works at the 50% of the maximum capacity, a random breakdown will introduce congestion in the network before splitting into two components. If the routing protocol is based in SP, the maximum capacity to avoid the congestion decreases to the 20% of the total. With higher values of the initial load, the system will fail dynamically before topologically. Finally, the robustness of random networks depends on the routing protocol. Using a SP, the communication process can operate up to the 80% of its capacity avoiding congestion.

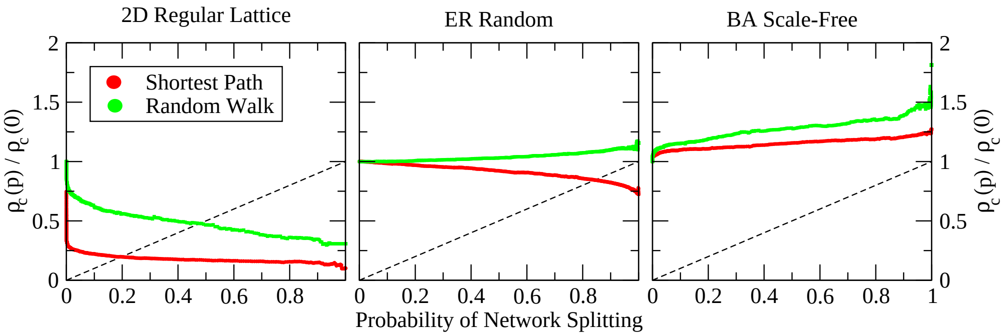


Figure 5.8. Dynamical robustness versus topological robustness. The dashed line delimits the probability of splitting the network. The crossing between the relative congestion ratio $\rho_c(p)/\rho_c(0)$ and the dashed line determine the change of the dominant effect between both processes. Before this point splitting dominates, beyond this point congestion dominates the communication process after random failures.

If we change the protocol to RW, we observe that the network improves its dynamical robustness, being very difficult to congest it via random failures.

5. Summary

In this last chapter we have studied the relationship between the random breakdown of a complex network and the changes on the congestion point of a communication processes. We have proved that this relationship is mainly governed by the algorithmic betweenness distribution. Moreover, we found that the centrality of the most important node in the communication process (the node with the highest betweenness) plays a crucial role in the changes of the onset of congestion. We presented an analytical expression for the behavior of the onset on congestion which is based on the amount of traffic that we remove from the node with maximum algorithmic betweenness, confirming its validity using different topologies and routing protocols. We also observed that if the breakdown modifies structural properties, the centrality of the nodes also changes, obtaining a different behavior of the congestion point.

The results provide some insight of the dynamical response of a network when there occur random failures. In other words, they give us an idea of the load a system can handle if we want to avoid the congestion, in case the network suffers random failures. These results highlight the necessity to include dynamical considerations in studies about resilience of complex networks. For instance, they can be used in the design of a dynamic communication process to guarantee the efficiency when some of the nodes have been removed.

Some interesting issues remain open for future studies. We expect that changing the topology or the routing protocols, we will be able to observe different slopes for $\rho_c(p)/\rho_c(0)$ which probably will be governed by the same constrains exposed in the work. Another appealing work derived from this problem is the analysis of the the congestion when the network undergoes an intentional attack.

Chapter 6

CONCLUSIONS AND PERSPECTIVES

1. Conclusions

In a time when large amounts of data about social, economical, technological, and biological systems are produced in a daily basis, complex networks have become a powerful tool to represent the structure of complex systems. The advances in complex networks theory have been geared towards the study of two main questions: what can we understand from a complex system by looking at its structure, and more importantly, what is the interplay between the topological and the dynamical properties of a complex system. The aim of this dissertation has been an attempt to provide new insights on both questions by analyzing two particular problem, the analysis of the community structure and the characterization of the dynamical properties of a communication process. Besides the particular summaries that are located at the end of each chapter, here we present the main conclusions that can be extracted from the obtained results.

- In the last few years, it has become clear that the detection of community structure of a complex network is key to characterize their internal organization. The identification of this intermediate scales of the system has enabled the scientific community to understand how the different components of a complex system assemble into coarser units, obtaining better insights about the dynamical behavior of these components.

  The problem of detecting communities has attracted the attention from scientists working in several fields, as the large number of efforts trying to quantify and detect this structure in the last five years reflect. In chapter two we have made a comprehensive comparative study of community detection methods in order to provide researchers with a guide on how to select the most appropriate method depending on the properties of the network or the computational resources available to extract a relevant structure.

Despite the fact that the analyzed methods follow completely different approaches, the accuracies obtained by most of them are similar when detecting well defined communities. Therefore, it seems that the problem of community detection provides a paradigmatic example of how a complex problem can be faced from a large number of points of view to reach the same conclusion. As we increase the "fuzziness" of the communities, we find that there is a trade-off between efficiency and accuracy. This is a common problem that appears when we want handle large amounts of data. In the particular case of community detection, the most accurate methods usually are not scalable so its use is limited to medium sized networks (up to 10,000 nodes). If we want to analyze larger networks, one needs faster alternatives, but their accuracy is lower.

- In chapter three we have presented an extremal optimization (EO) based method to detect community structure as an alternative to those available in the literature. The aim of proposing a new method is two-fold: first, to minimize the trade-off of efficiency or accuracy that we commented in chapter two; and second, to provide a novel approach to the problem based on a different type of heuristics.

  Even though the extremal optimization is not as popular as other classical heuristics such as simulated annealing or genetic algorithms, it yields very good results when applied to classical optimization problems. In addition, since behind EO there is an evolutionary process where nodes self-organize until they reach a stable configuration that gives the community structure, EO can also be used to understand the process behind the formation of the communities. When comparing the results of our method with those presented in the benchmark of chapter two, we find that our method is among the most accurate, and is also able to perform the community analysis in a short amount of time. For these reasons, results obtained with our algorithm have been used as a comparative reference in posterior detection methods.

  In this chapter we have also performed a deeper analysis of the modularity formulation. We redefined this measure in terms of weighted and directed versions, and we have uncovered the contribution to the modularity equation of particular configurations of nodes. Using this knowledge and the flexibility of the EO algorithm, we have been able to introduce small modifications that go a step further in the detection process. On one hand, with few algorithmic details, we have improved the accuracy of recursive algorithms up to a 20%. On the other hand, we have proposed a method on how to reduce the size of a network up to a 40% without altering the maximal modularity configuration, allowing for even faster and more accurate community analysis.

- In the last two chapters we have studied of some dynamical properties of communication processes in complex networks. We believe that opening or reinforcing research lines devoted to the study of global scale properties of dynamic process is mandatory. When scientists model the topology of a specific complex network (e.g. the Internet), they measure the structural properties (e.g. they find a scale-free degree distribution) and reproduce them in the model, obtaining a simple but accurate description of reality. However, we cannot follow the same steps to model the dynamics because we lack the equivalent set of measurements that we use for the topology.

  In this context, we have studied the fluctuations of the traffic on complex networks in order to provide a large-scale dynamical characterization of the traffic flow. The idea is that there are a large number of real complex systems that show a scaling relation between the average flux and the variability of this flux. The understanding of this scaling relation will help us design better traffic models.

  It seems true that the values that the scaling exponents $\alpha = 1/2$ and $\alpha = 1$ proposed in (de Menezes and Barabási, 2004a, de Menezes and Barabási, 2004b) are important, since act as a bound of the scaling observed in real systems. However, in chapter four we show that these values are not universal, and that between the two exponents there is a wide range of possible values that appear by tuning the parameters that control the communication process. The analysis of a simple traffic model based on a Poisson queueing system reveals three mechanisms that give rise a transition: how we perform the sampling process, the time that the packets stay on the network, and the introduction of congestion into the networks.

  We corroborate the existence of intermediate exponents in real systems by looking at the fluctuations of Internet traffic, which can be characterized by a a scaling exponent $\alpha \sim 0.75$. Our results are in agreement with other studies of fluctuation scaling in complex systems that display different scalings in the range $[1/2, 1]$ (Eisler et al., 2007).

- The capability to maintain the communication between nodes when some of them fail is another interesting property worthy analyzing. We believe that it is more important to observe the incidence of the node failures on the dynamical processes supported by the network than the effect of the failures on the topology. In chapter five, we have defined the dynamical robustness of a communication process as the ability of the traffic to avoid congestion when we remove a fraction of nodes. Using the same communication model of chapter four, we have provided new insights on how the onset of congestion behaves when there is a random node removal for different routing protocols over different network topologies.

We have performed a theoretical study of the problem and we have quantified the changes produced by the random breakdown by analyzing the centrality of the most important node in the communication process, and the average number of jumps that the packets perform before exiting the network. It is also worth noting that when we consider a protocol in which the nodes do not have global knowledge of the structure, we find that changes on the average path length of the system control the different behaviors that appear in the simulations.

2. Perspectives

Here we resume some research lines that provide a logical continuation of the work presented in this thesis.

- There are still open questions regarding the internal organization of complex networks. The first one is in the definition of what is a community. It seems that the modularity has been accepted as the 'de facto' measure to quantify a given structure. However, some studies have also pointed out the weak points of this measure, particularly the limitation to identify certain structures that can be relevant in the dynamical processes. This opens the door to alternative quantitative measures more suitable to capture dynamically relevant structures.

  The second open question is the study of overlapping and hierarchical communities. New methods have been proposed to identify all the mesoscopic scales of the system and communities that share certain nodes, but again we find the same problems of which is the method that we should use to analyze one network. In this case an extension of the benchmarks and tools presented in chapter 2 should be developed to help the scientists decide.

  Finally, the last —and perhaps the most interesting— open question refers to the study of community dynamics. The analysis of the processes responsible of the formation and the evolution of the community structure will provide new insights about social and/or economical dynamics. The availability of data will unable us to track changes on the community structure with time, providing us with a clearer picture on how the nodes self-organize into these groups.

- The extremal optimization algorithm to detect community structure presents nowadays a stable solution. Apart from minor technical improvements, the method can be extended in at least three directions that will increase the number of problems where we could apply our method. First, we can take advantage of the flexibility of the algorithm and change the cost function to look for groups with other properties (such as node similarity as suggested in (Newman and Leicht, 2007)), obtaining partitions of nodes that share these properties. This task is not trivial, since we first need to define the function that we want to optimize and then define the individual contribution of the nodes to this function.

Second, we can extend the method by including the mesoscopic analysis introduced in (Arenas et al., 2007). Combining the accuracy of our algorithm with this method will provide a more comprehensive picture of the different organizational levels in a complex network.

Third, we can use our method to solve the community detection problem in a probabilistic fashion, that is, we can group the nodes considering the probability of them belonging to the same community, instead of giving only one configuration that corresponds to the maximum modularity. Since our method is not deterministic, by repeating the analysis of a network several times we usually obtain different configurations which have small structural changes. By performing an statistical analysis of all the output configurations, we can extract the probability that nodes belong to each community. A similar approach has been already successfully applied in (Sales-Pardo et al., 2007).

- The study of the scaling of fluctuations is in a preliminary stage, thus there is a large number of questions unanswered. In order to improve our understanding of the problem presented here, a possible alternative is the exploration of other possible transitions of the scaling exponent by using a a more realistic traffic model. To increase the complexity of the model we can also use other routing protocols (such as shortest-path based), or change the behavior of the queueing system. We expect that the introduction of these changes will give rise to new transitions of the exponent that could help us understand one of the key questions of the problem: what are the reasons that determine the exponent for a particular real system?. The answer to the question is not trivial since there are many explanations for the existence of a given value of α. A good way to understand the exponent should be the study of the different parameters of the system and, based on the model, look for which ones can give a coherent explanation of the fluctuation.

  The communication model used in the chapter is unable to reproduce the self-similarity of traffic in time observed in some real systems like the Internet. Therefore, a natural extension of the work is to introduce traffic with long-range dependencies and burstiness, and look at the behavior of the system. Moreover, the use of this type of traffic opens the door to another research line: the study of the relationship between the scaling exponents and the Hurst exponent that governs the self-similarity of the traffic.

- The study of the dynamical robustness is also in its preliminary stages. The first problem that we encountered by using this methodology relies in the amount of computing resources needed to perform a simulation of the system and obtain the values of the onset of congestion. An alternative way to obtain the value of ρ_c is the use of the formulation proposed in (Guimerà et al., 2002b). However, the adaptation of the formulation is not trivial, since it requires the inversion of a big matrix which also uses a large number of computational resources. Using techniques such as the LU decomposition or the Cholesky decomposition (Golub and van Loan, 1996), we expect to be able to reduce substantially the time needed to reduce the matrices and as a result, the time needed to analyze the changes on the onset of congestion.

  Another possibility that would be extremely interesting is to test the dynamical robustness of real network topologies, to be able to corroborate from an empirical point of view the results obtained in the simulations of the work. And finally, another possible extension is the analysis of what happens with the onset of congestion when the network suffers directed attacks against the most important nodes, such as the most connected or the most central in the communication process.

Appendix A
Evolution of the Internet Autonomous System Topology

In the introduction of this work we have explained the reasons why the scientific community is still looking for a detailed map of the Internet's topology. We also have explained how Internet mapping projects try to solve this problem, collecting information using passive (BGP tables) and active (traceroute probes) mechanisms, filtering the gathered information and creating punctual snapshots of the Autonomous Systems (AS) and Internet Router (IR) levels. These maps are published online and research groups are using them to test new models, theories or protocols.

A first look to some of the available online maps shows a topology with a large degree of heterogeneity, independently on the source used to gather the data. This observation was confirmed empirically by Faloutsos *et al.* when they found that the degree distribution of the Internet obeys a power-law (Faloutsos et al., 1999). Since then, scientists have used the set of the statistical tools and measurements described in the first chapter when they want to perform a large scale characterization of the Internet network (Vázquez et al., 2002a).

The evolution of the Internet modeling has been directly related to the observation of these statistical measurements. Every time that a new model is developed, the authors take one or various topology datasets, measure some of their large-scale properties, and compare them against their model to validate it. This validation process has been criticized since the measures obtained from the datasets used in the comparison between model and reality have some potential flaws. First, some recent publications argue that the obtained statistical measurements can be biased by the method used to extract the topology (Lakhina et al., 2002, Clauset and Moore, 2005), leading to an uncomplete view of the network that cannot reflect the real structure behind. For instance, it has been proved how a bad sampling of a homogeneous graph can make us believe that the inferred network has a scale-free distribution (Dall'Asta et al., 2005). Computer scientists are working hard to solve the sampling problem, creating more efficient reverse engineering techniques or increasing the number of nodes used in the active mechanisms that extract the topology information (Donnet et al., 2005). And second, another group of critics argue that the Internet is a network that has been continuously growing and changing at all the three scales (users, routers and Autonomous Systems) in the last 20 years. Every month new nodes appear and some of them disappear, turning the evolution of the network into what is known as a "birth-death process" that can change some of the structural properties of the network. Therefore, one model that is capable of reproducing a concrete snapshot of the network could not be valid a few months before.

In this appendix we present a brief analysis of the evolution of the statistical properties of the Internet AS maps to measure the stability of internet measurements, extending the previous works of Vázquez et al. (Pastor-Satorras et al., 2001, Vázquez et al., 2002b, Vázquez et al., 2002a). The presented results can be used as a reference in future modeling of the Internet Autonomous System topology.

1. Mapping the Autonomous System topology

An Autonomous System refers to a set of routers that are under a single technical administration, where communications between routers within the AS are controlled by an interior protocol and communications to other ASs by an exterior protocol (usually BGP). From a more restrictive point of view, the RFC documentation[1] fixes as a condition to be an Autonomous System that the entity should have a single and clearly defined internal routing policy (Hawkinson and Bates, 1996).

An Autonomous System usually refers to administrations or Internet Service Providers that comply with the previous conditions. When two ASs want to exchange traffic with each other, they need to establish a physical connection between them. But behind the creation of a new connection there are a large number of issues that need to be negotiated by both parts, e.g. who is going to pay for the infrastructure or how much traffic is allowed to travel in each direction. Therefore, it is difficult to classify this network as technological, since each link involves some kind of agreement between the two involved parties. This provides another point of view of the network, the social network of agreements between all the entities that can give access to the Internet around the world.

[1] The Request For Comments (RFCs) documentation is considered the official collection of all the designs and guidelines realted with Internet

Internet AS maps can not be directly obtained by simply looking at the network infrastructure. Data collected from BGP tables or from traceroute-like techniques gives only the connectivity between router interfaces. Since the autonomous system represents a coarse-grained view of the network, there is a necessary abstraction process before obtain the representation of the network at the AS level. The first step towards obtaining the AS map begins with the extraction of the connectivity graph between routers. Then, routers are grouped using the information regarding their ownership, creating a new network where each node represents an AS. The links between two ASs are added if there is at least one physical connection between the routers belonging to them. The resulting new network goes through a filtering process to correct duplicate entries and to validate the results (Dimitropoulos et al., 2007), and finally the resulting map is published in the online repositories.

The AS topologies used in the analysis have been gathered from some of the data sources most frequently used by the research community. Some topology mapping projects publish new datasets regularly (usually one or two times a week) that contain already preprocessed information about the relationship between ASs. The three sources of information that we have used are the following:

- The first group of data comes from the National Laboratory for Applied Network Research (NLANR), who generated AS maps derived from the BGP routing tables gathered by the Routeviews project from 8 November 1997 to 2 February 2000. The datasets are freely available for download at http://moat.nlanr.net/Routing/rawdata/.

- The second group of data has been obtained from the Cooperative Association for Internet Data Analysis (CAIDA), who since 2004 has been publishing the adjacency matrix of the Internet AS-level graph. This data is also inferred from the BGP tables of Routeviews project, but in this case they also add extra information about the type of link between two ASs (Customer-Provider, Peer-to-Peer, etc..). The datasets are freely available for download at http://as-rank.caida.org/data/.

- The third group of data is a combination of the data from the previous group with the measurements of the Skitter project, which is also managed by CAIDA. The Skitter project uses traceroute probes to discover hidden paths between the ASs present in the BGP tables. Since it is a complementary dataset to the previous one, we have combined the information of the links of both sources to present a more detailed map of the Internet AS-level. Skitter data is freely available for download at http://sk-aslinks.caida.org/.

We have downloaded their datasets and created monthly snapshots of the AS maps by combining all the nodes and links that appear at least twice in the

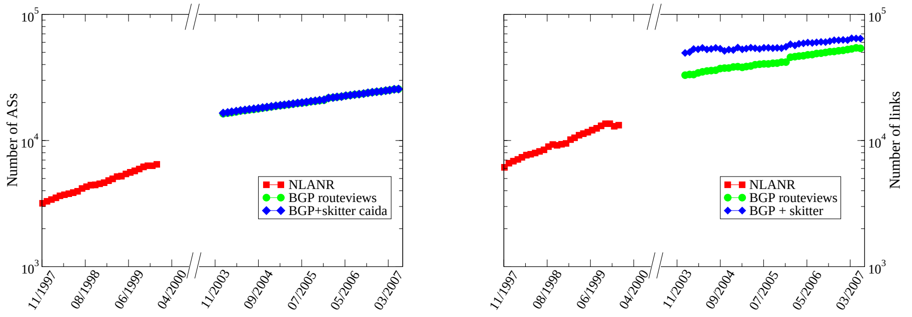


Figure A.1. Evolution of the number of nodes (a) and of links (b) of the Autonomous Systems topology maps. The nodes and links are shown in a logarithmic scale.

networks published in a given month to discard instabilities of the data that can give rise to false links. Note that there is a temporal gap between the two datasets, since in this period there was no group publishing data about AS relationships. Between 2001 and 2004 there are available BGP information in the Routeviews project, but no project was creating the AS relationship maps. We guess that one possible reason to explain this lack of information comes from the “Dot-Com bubble” crash of 2001, which decreased momentarily the interest and funding for Internet related projects. The interest in Internet measurements (and the AS relationships mapping) raised again with the establishment of CAIDA as a worldwide internet observatory, who has been releasing more detailed datasets since the beginning of 2004.

2. Analyzing the evolution of the AS maps

To study the evolution of the structure of the AS network, we have analyzed the changes of some of the most common measurements used in complex networks literature: the size of the system, degree based properties, small-worldliness, and the hierarchical organization of the network.

First, we have studied the change on the size of the Internet AS map. In figure A.1, we observe that there is an exponential growth of the number of nodes and links in the three datasets. Keep in mind that this growth refers to the number of organizations that provide access to the network, and not to the number of people that uses its services.

To quantify the growth rates we have used the same formulation presented in (Serrano et al., 2005). We characterize the exponential growing rates as $E(t) \sim E_0 e^{\alpha t}$ for the number of links and $V(t) \sim V_0 e^{\beta t}$ for the number of nodes, where t is the number of months that have passed since the first measurement of each dataset. The exponents found in the NLANR datasets are $\alpha \sim 0.03$ and $\beta \sim 0.027$, in concordance with those obtained in Serrano *et al.*. In the CAIDA datasets the exponents are $\alpha \sim 0.012$ and $\beta \sim 0.011$

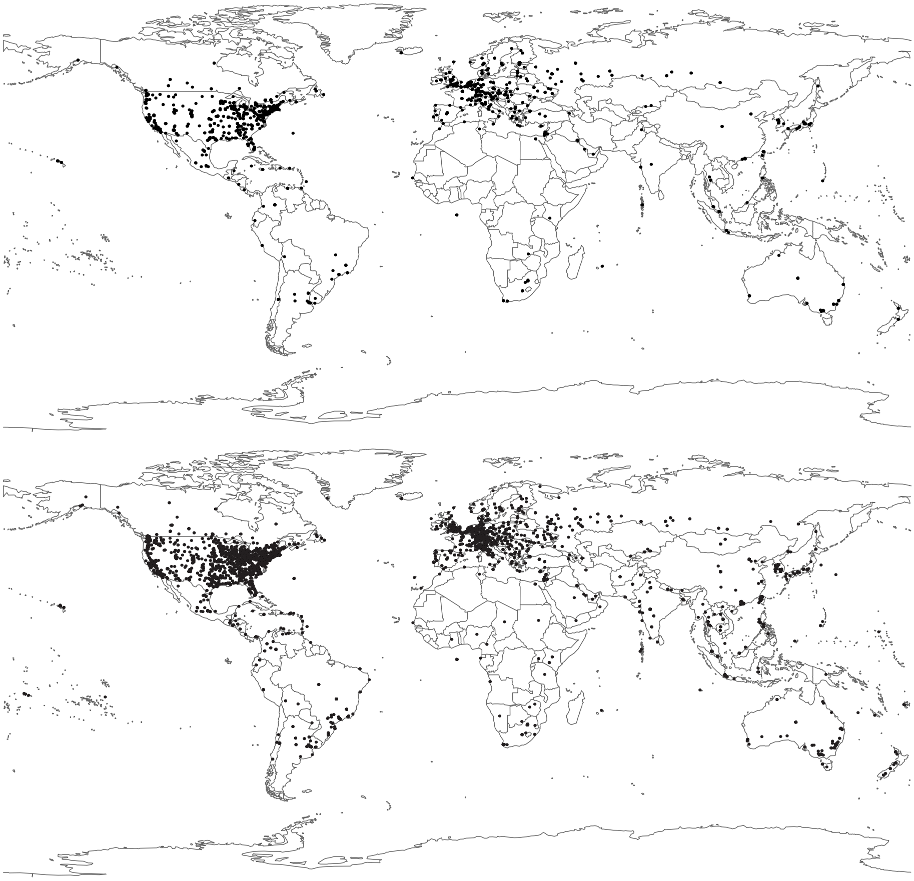


Figure A.2. Geographical distribution of the Internet AS nodes in the inferred maps by NLANR in 1997 (c) and by CAIDA in 2007 (d). We see that the growth of the AS network has been centered in the most developed countries.

without the Skitter data and $\alpha \sim 0.005$ and $\beta \sim 0.011$ with the Skitter extra information.

The number of nodes and links is still nowadays growing exponentially, but it seems that the growing rate has decreased by half. One of the most plausible explanations for this fact is that there is a saturation of the market of the Autonomous Systems in the most developed countries, which is decreasing the probability that new ASs enter the market. To evaluate the plausibility of this hypothesis we have plotted the geographical position of the Internet ASs at two different times. The results are presented in figure A.2. As expected, the growth of the system has been centered in two main areas: the United States and Europe, where it covers almost all the high populated areas and therefore seems that there could be some saturation of the market in these areas. The fractal distribution observed in this two plots has been compared with the distribution of the world population (Yook et al., 2002), and it seems that it plays an important role in the growth rules of the Internet topology (Serrano et al., 2005, Serrano et al., 2006).

Finally, also note the difference in the number of links between the two CAIDA datasets in figure A.1 right. The extra information provided by Skitter reveals the existence of the previously described Internet ’dark-matter’, a large number of links (between 10 and 15 thousand links) that are not detected if we create a map only using the BGP routing tables. The number of nodes in both datasets are the same, since Skitter only uses the nodes that appear in the BGP tables to test the existence of the hidden links between them.

Maximum degree, average degree, degree distribution and correlations

One of the most frequently used group of metrics that have been used to characterize the Internet large-scale topology —which does not mean that they are the best ones, as explained in chapter one— have been based on the number of connections of a nodes. In particular, we have analyzed four interesting metrics of the degree: the average node degree, the maximum degree, the degree distribution and the degree-degree correlations.

Looking at the average degree of the nodes in figure A.3.a we observe that the AS level is a very sparse network, even if we add the extra links from the Skitter. The value of the average degree has been almost constant in the last 10 years, fluctuating between 4 and 4.5. Adding more links from Skitter obviously increases the average value, but the difference is very small; seems that the if we can discover all the links, the average of connections per node will be around 5 and 6. On the contrary, looking at the maximum degree of the AS network in figure A.3.b, we find that this metric was growing linearly in the NLANR maps, but seems that in the last three years has reached an stationary value around 3000 links (a little less if we do not add the Skitter ones).

The connectivity distribution of Internet is known to display a scale-free distribution. We have measured the accumulated distribution of degrees for both CAIDA datasets of May 2007 in figure A.3.c, and we find a value around -1.1 in agreement with the values obtained in other analysis (Vázquez et al., 2002b). In figure A.3.d we plot the evolution of the exponent in the different datasets. We observe that the value is almost stationary in time, and that is not affected by the introduction of additional links from Skitter. The stability of the exponent can be explained if the growing process behind the addition of nodes is based in the preferential attachment mechanism (Barabási and Albert, 1999). As Barabási *et al.* pointed out, when a network created with preferential attachment reaches a steady state, the exponent of the degree distribution is invariant to changes on the size of the network.

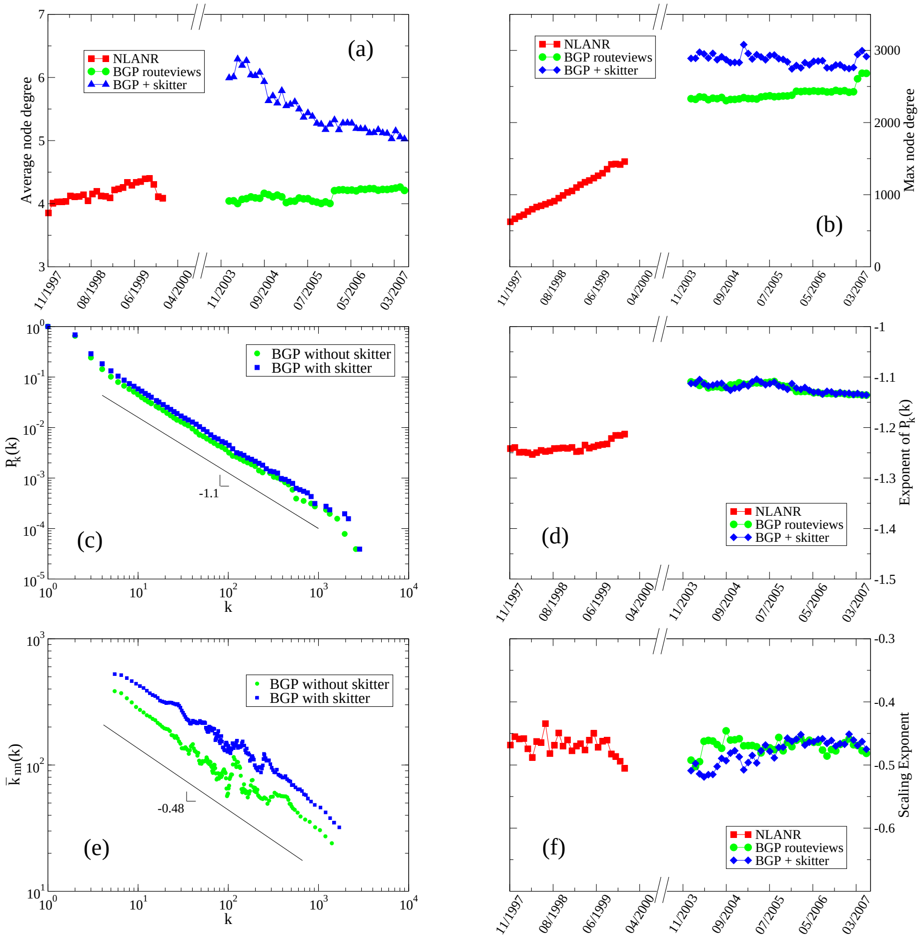


Figure A.3. Evolution of the maximum degree (a) and average degree (b) of the analyzed AS topology maps. Accumulated degree distribution of the two CAIDA snapshots of May 2007 (c) and evolution of the exponent of the accumulated degree distribution (d). Average degree of the nearest neighbors as a function of k of the two CAIDA snapshots of May 2007 (e) and evolution of the scaling exponent between the degree and the average degree of the nearest neighbors (f).

Another measure related with the node degrees is the degree-degree correlation function. Figure A.3.e displays the average degree of the nearest neighbors of a vertex $\overline{k}_{nn}$ as a function of the node degree k, where we observe a scaling relationship with an exponent ~ -0.48. The interpretation of this observation is that the AS-map display disassortative mixing, where high degree nodes are on average mostly connected to nodes with a smaller degree. This is not a surprising result, since many technological based networks (including the IR map) also display a clear disassortative mixing.

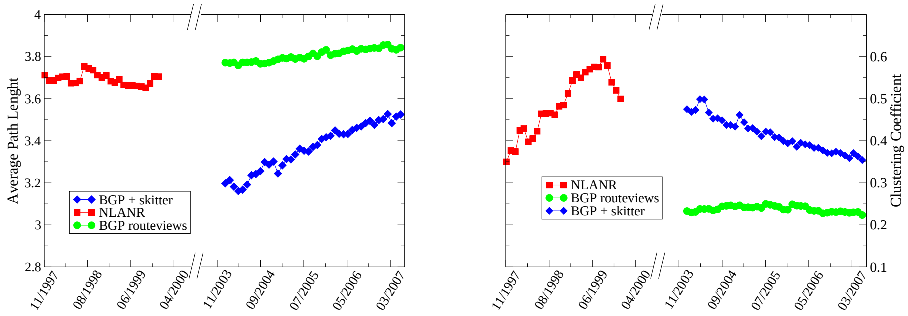


Figure A.4. Evolution average shortest path length (left) and of the clustering coefficient (right) of the AS topology maps.

On figure A.3.f we show the evolution of the scaling exponent of the $\overline{k}_{nn}(k)$ distribution. This exponent proves a very stable metric with a value around -0.47, that has not been affected by the evolution and by adding more links into the network.

Small-world properties: clustering and average path length

Popular culture affirms that one of the most interesting advances of the Internet is that it has given us the feeling than the world is smaller than ever before. However, does the internal structure of the network display the characteristics of a small-world network? A few studies have pointed out that the AS-level topology is indeed a small-world, with a very short average path length between its nodes and a relatively high clustering coefficient (Bu and Towsley, 2002). This is not surprising since the AS map is composed of many local ASs highly connected between them (that gives rise to a high clustering coefficient), which are connected by large ASs that act as backbones of the network, as depicted in the Transit-Stub model (Zegura et al., 1996). The existence of this small-worldliness is also important for dynamical processes, since seems that plays a key role on the efficiency of the Internet on delivering information (Latora and Marchiori, 2001).

Figure A.4 shows the evolution of the average path length and the clustering coefficient metrics of the analyzed maps. On one hand, we find that the average path length on the BGP maps has been almost constant in the last seven years, increasing only around 1% when the size of the network has increased 150%. The main reason behind this behavior is the scale-free nature of the AS maps; as pointed out by Cohen et al., scale-free networks are “Ultra-small” (Cohen and Havlin, 2003) because their diameter scales as the size of the system as $\langle D \rangle \sim \log \log N$. Since the AS maps display a well-defined scale-free degree distribution, we expect that adding new nodes will not change the distances on

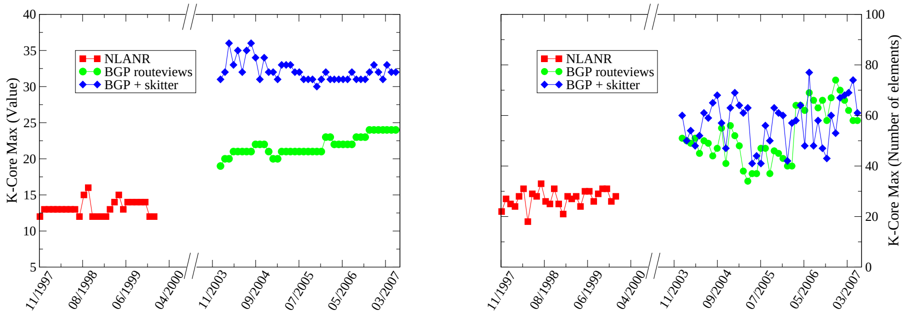


Figure A.5. Evolution of the value (left) and size (right) of the maximum k-core. The fluctuations are a consequence of the sampling process used to infer the maps by the mapping projects.

the network. The inclusion of the Skitter links decreases the average distance, as we provide alternate paths to reach other nodes of the network while we do not change the number of nodes.

On the other hand, we observe that the value of the clustering coefficient has been changing continuously. There was an increasing trend in the first months of the NLANR, but seems that now the value is going to a stable value around 0.2. The inclusion of more links from Skitter while maintaining the number of nodes increases the clustering coefficient of the network up to 0.35, since the missing links are mostly redundant connections between the most connected nodes (Cohen and Raz, 2006). This is one of the main reasons why this links are not detected with traditional techniques that only look at the main paths.

Hierarchical structure of internet: k-cores

Another important feature of the Internet AS structure is that displays a well-defined hierarchical structure. One possibility to measure this hierarchical organization is using the k-core decomposition. The k-core decomposition consists of a recursive pruning process of the least connected vertices, obtaining the most central core of one network and uncovering its hierarchical organization (Seidman, 1983, Bollobas, 1983).

The size and degree of the largest k-core gives information about the robustness of the network and its potential efficiency. In first place, the existence of a big k-core means that in the center of the network there is a big number of nodes interconnected, which decreases the probability of breaking the network. In second place, the redundancy of the links helps the redistribution of the traffic flow among a greater number of paths, providing higher efficiency on delivering the information between pairs of nodes (Alvarez-Hamelin et al., 2007).

In the last two plots we analyze the evolution of this two metrics of the k-core decomposition, the degree and the size of the largest k-core. In figure

A.5 we show that both values have been increasing slowly with the Internet evolution. The addition of the Skitter links increases substantially the degree of the k-core nodes, reinforcing the fact that the missing links of the BGP tables correspond mainly to recursive connections between the most central nodes (Cohen and Raz, 2006).

The values observed in this two figures show that the Internet has a large core of around 65 nodes that are highly connected between them (each node is connected to half of the others). However, these values are not in agreement with the presented in the Internet medusa model (Carmi et al., 2007), where they found a core of 100 nodes. The most probable reason is that the authors of this model use the AS maps published by the DIMES project, which include different information about the missing links, and therefore they observe a different snapshot of the AS maps.

3. Summary

In this appendix we have shown how the internal structure of the Internet AS maps has been evolving in the last seven years. We have studied the temporal changes of an extensive set of topological characteristics, including average properties and exponents of scale-free distributions. We have analyzed the AS maps published on two different time frames: the data collected from NLANR between 1997 to 2000 displays how some properties are changing, probably because since the core of Internet was still in formation; in both AS maps collected from CAIDA between 2004 and 2007 we find that the evolution of its structure is getting into a mature state, with an internal structure that is insensitive against the addition and removal of nodes.

There is an important open question regarding these results: Are the changes in the AS maps reflecting the Internet evolution or they are a consequence of the increasing efficiency of Internet discovery tools? On one hand, the discovery tools are providing more detailed information about the topology, as we have seen when we added the Skitter links to the BGP tables. On the other hand, no one doubts that Internet is still growing at all its levels, and therefore there is more information to discover every month. Seems that to answer this questions we will need to until we will have a complete topology map, and then we will be able to identify exactly how much of the evolution is by the improvement of the tools and how much is due to the real growth of the Internet.

Appendix B
Relationship between directed and undirected modularities

Let us suppose that w_{ij} are the weights of a directed weighted network, and that we define its corresponding symmetrized (undirected) network by adding the weights matrix to its transpose:

$$\bar{w}_{ij} = w_{ij} + w_{ji}\,,\ \forall i,j\,. \tag{B.1}$$

The strengths of this undirected network are

$$\bar{w}_i = w_i^{\text{out}} + w_i^{\text{in}}\,, \tag{B.2}$$

and the total strength is

$$2\bar{w} = 4w\,. \tag{B.3}$$

The modularity Q_D of the directed network is invariant under transposition of the weights matrix since the input (output) strengths of the transposed network are equal to the output (input) strengths of the original one:

$$\begin{aligned} Q_D &= \frac{1}{2w}\sum_i\sum_j\left(w_{ij} - \frac{w_i^{\text{out}}w_j^{\text{in}}}{2w}\right)\delta(C_i,C_j) \\ &= \frac{1}{2w}\sum_j\sum_i\left(w_{ji} - \frac{w_j^{\text{out}}w_i^{\text{in}}}{2w}\right)\delta(C_j,C_i) \\ &= \frac{1}{2w}\sum_i\sum_j\left(w_{ji} - \frac{w_i^{\text{in}}w_j^{\text{out}}}{2w}\right)\delta(C_i,C_j)\,. \end{aligned} \tag{B.4}$$

The relationship between the modularity Q_D of the directed network and the modularity Q_S of its symmetrization is obtained by simple calculations:

$$\begin{aligned}
Q_S &= \frac{1}{2\bar{w}} \sum_i \sum_j \left(\bar{w}_{ij} - \frac{\bar{w}_i \bar{w}_j}{2\bar{w}} \right) \delta(C_i, C_j) \\
&= \frac{1}{4w} \sum_i \sum_j \left(w_{ij} + w_{ji} - \frac{(w_i^{\text{out}} + w_i^{\text{in}})(w_j^{\text{out}} + w_j^{\text{in}})}{4w} \right) \delta(C_i, C_j) \\
&= \frac{1}{4w} \sum_i \sum_j \left[\left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) + \left(w_{ji} - \frac{w_i^{\text{in}} w_j^{\text{out}}}{2w} \right) \right] \delta(C_i, C_j) \\
&\quad - \frac{1}{(4w)^2} \sum_i \sum_j (w_i^{\text{out}} - w_i^{\text{in}})(w_j^{\text{out}} - w_j^{\text{in}}) \delta(C_i, C_j) \\
&= Q_D - \frac{1}{(4w)^2} \sum_i \sum_j (w_i^{\text{out}} - w_i^{\text{in}})(w_j^{\text{out}} - w_j^{\text{in}}) \delta(C_i, C_j) \,. \qquad \text{(B.5)}
\end{aligned}$$

This result can also be expressed as a communities sum:

$$Q_S = Q_D - \frac{1}{(4w)^2} \sum_r \left(\sum_i (w_i^{\text{out}} - w_i^{\text{in}}) \delta(C_i, r) \right)^2 . \qquad \text{(B.6)}$$

The contribution of the links to the input and output strengths cancel if they fall within the communities. Therefore, if most links do not cross the boundaries of the communities, it follows that $Q_S \approx Q_D$ even if the network is highly asymmetric.

Appendix C
Analytic network reduction preserving modularity

As we have introduced in chapter three, one can reduce the size of a network grouping nodes while preserving the modularity. In this appendix we give the analytical proof of the modularity preservation, and its application to two different size reductions of weighted networks. This two reductions can be applied in both undirected and directed networks.

1. Size reduction preserving modularity

The main property of the reduced network is the preservation of modularity, i.e. the modularity of any partition of the reduced graph is equal to the modularity of its corresponding partition of the original network.

More precisely, let $C' : \{1, \ldots, N'\} \longrightarrow \{1, \ldots, M\}$ be a partition in M clusters of the reduced network G'. Its corresponding partition $C : \{1, \ldots, N\} \longrightarrow \{1, \ldots, M\}$ of the original graph is given by the composition of the reducing function R with the partition C', i.e. $C = C' \circ R$. Therefore, the statement of the previous paragraph becomes

$$Q'(C') = Q(C)\,. \tag{C.1}$$

The proof is straightforward:

$$Q'(C') = \frac{1}{2w'} \sum_r \sum_s \left(w'_{rs} - \frac{{w'_r}^{\text{out}} {w'_s}^{\text{in}}}{2w'} \right) \delta(C'_r, C'_s)$$

$$
\begin{aligned}
Q'(C') &= \frac{1}{2w}\sum_r\sum_s\left(\sum_i\sum_j w_{ij}\delta(R_i,r)\delta(R_j,s)\right.\\
&\quad \left. -\frac{1}{2w}\sum_i w_i{}^{\text{out}}\delta(R_i,r)\sum_j w_j{}^{\text{in}}\delta(R_j,s)\right)\delta(C'_r,C'_s)\\
&= \frac{1}{2w}\sum_i\sum_j\left(w_{ij}-\frac{w_i{}^{\text{out}}w_j{}^{\text{in}}}{2w}\right)\sum_r\sum_s\delta(R_i,r)\delta(R_j,s)\delta(C'_r,C'_s)\\
&= \frac{1}{2w}\sum_i\sum_j\left(w_{ij}-\frac{w_i{}^{\text{out}}w_j{}^{\text{in}}}{2w}\right)\delta(C'_{R_i},C'_{R_j})\\
&= \frac{1}{2w}\sum_i\sum_j\left(w_{ij}-\frac{w_i{}^{\text{out}}w_j{}^{\text{in}}}{2w}\right)\delta(C_i,C_j)\\
&= Q(C) \qquad \text{(C.2)}
\end{aligned}
$$

2. Reductions for undirected networks

The modularity of an undirected network may be written as

$$Q=\sum_i q_i\,, \qquad \text{(C.3)}$$

where

$$q_i=\frac{1}{2w}\sum_j\left(w_{ij}-\frac{w_iw_j}{2w}\right)\delta(C_i,C_j) \qquad \text{(C.4)}$$

is the contribution to modularity of the i-th node. If we allow this node to change of community, the value of C_i becomes a parameter, and therefore it is useful to define

$$q_{i,r}=\frac{1}{2w}\sum_j\left(w_{ij}-\frac{w_iw_j}{2w}\right)\delta(C_j,r)\,,\quad q_i=q_{i,C_i}\,, \qquad \text{(C.5)}$$

which accounts for the contribution of the i-th node to modularity if it were in community r. The separation of the self-loop term, which does not depend on which community node i belongs to, yields to the definition of

$$\tilde{q}_{i,r}=\frac{1}{2w}\sum_{j(\neq i)}\left(w_{ij}-\frac{w_iw_j}{2w}\right)\delta(C_j,r)\,,\quad \tilde{q}_i=\tilde{q}_{i,C_i} \qquad \text{(C.6)}$$

and

$$\tilde{Q} = \sum_i \tilde{q}_i = \frac{1}{2w} \sum_i \sum_{j(\neq i)} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_j, r) \,, \tag{C.7}$$

satisfying

$$q_{i,r} = \tilde{q}_{i,r} + \frac{1}{2w} \left(w_{ii} - \frac{w_i^2}{2w} \right) \tag{C.8}$$

and

$$Q = \tilde{Q} + \frac{1}{2w} \sum_i \left(w_{ii} - \frac{w_i^2}{2w} \right) \,. \tag{C.9}$$

The role of these individual node contributions to modularity becomes evident in the expression of the change of modularity when node i goes from community r to community s:

$$\Delta Q = 2(\tilde{q}_{i,s} - \tilde{q}_{i,r}) \,. \tag{C.10}$$

As a particular case, a node that forms its own community, i.e. an isolated node i, which moves to any community s produces a change in modularity

$$\Delta Q = 2\tilde{q}_{i,s} \,. \tag{C.11}$$

Therefore, if there exists a community s for which $\tilde{q}_{i,s} > 0$, node i cannot be isolated in the partition of optimal modularity. This existence is easily proved by considering the sum of $\tilde{q}_{i,r}$ for all communities:

$$\begin{aligned}
\sum_r \tilde{q}_{i,r} &= \frac{1}{2w} \sum_{j(\neq i)} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \sum_r \delta(C_j, r) \\
&= \frac{1}{2w} \sum_{j(\neq i)} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \\
&= -\frac{1}{2w} \left(w_{ii} - \frac{w_i^2}{2w} \right) \,.
\end{aligned} \tag{C.12}$$

where we have made use of the definitions of strength w_i and total strength $2w$ for the simplification of the expression. Thus,

$$\text{if } w_{ii} \leq \frac{w_i^2}{2w} \;\;\Rightarrow\;\; \sum_r \tilde{q}_{i,r} \geq 0 \;\;\Rightarrow\;\; \exists s : \tilde{q}_{i,s} \geq 0 \,, \tag{C.13}$$

completing the proof that there are no isolated nodes in the configuration which maximizes modularity, unless they have a big enough self-loop.

3. Reductions for directed networks

The treatment of directed networks requires the distinction between the nodes' output and input contributions to modularity. We have proved in the previous appendix that the modularity is invariant under the transposition of the weights matrix:

$$Q = \sum_i q_i^{\text{out}} = \sum_j q_j^{\text{in}} \,, \tag{C.14}$$

where

$$q_{i,r}^{\text{out}} = \frac{1}{2w} \sum_j \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_j, r) \,, \quad q_i^{\text{out}} = q_{i,C_i}^{\text{out}} \,, \tag{C.15}$$

$$q_{j,r}^{\text{in}} = \frac{1}{2w} \sum_i \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_i, r) \,, \quad q_j^{\text{in}} = q_{j,C_j}^{\text{in}} \,. \tag{C.16}$$

The process of separating the self-loop term follows the same pattern than for undirected networks:

$$\tilde{q}_{i,r}^{\text{out}} = \frac{1}{2w} \sum_{j(\neq i)} \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_j, r) \,, \quad \tilde{q}_i^{\text{out}} = \tilde{q}_{i,C_i}^{\text{out}} \,, \tag{C.17}$$

$$\tilde{q}_{j,r}^{\text{in}} = \frac{1}{2w} \sum_{i(\neq j)} \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_i, r) \,, \quad \tilde{q}_j^{\text{in}} = \tilde{q}_{j,C_j}^{\text{in}} \,, \tag{C.18}$$

and

$$\tilde{Q} = \sum_i \tilde{q}_i^{\text{out}} = \sum_j \tilde{q}_j^{\text{in}} \,, \tag{C.19}$$

satisfying

$$q_{i,r}^{\text{out}} = \tilde{q}_{i,r}^{\text{out}} + \frac{1}{2w} \left(w_{ii} - \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \right) \,, \tag{C.20}$$

$$q_{j,r}^{\text{in}} = \tilde{q}_{j,r}^{\text{in}} + \frac{1}{2w} \left(w_{jj} - \frac{w_j^{\text{out}} w_j^{\text{in}}}{2w} \right) \,, \tag{C.21}$$

and

$$Q = \tilde{Q} + \frac{1}{2w} \sum_i \left(w_{ii} - \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \right) . \tag{C.22}$$

With these definitions at hand, the change of modularity when node i goes from community r to community s becomes

$$\Delta Q = (\tilde{q}_{i,s}^{\text{out}} + \tilde{q}_{i,s}^{\text{in}}) - (\tilde{q}_{i,r}^{\text{out}} + \tilde{q}_{i,r}^{\text{in}}) , \tag{C.23}$$

and the change when an isolated node i moves to any community s is

$$\Delta Q = \tilde{q}_{i,s}^{\text{out}} + \tilde{q}_{i,s}^{\text{in}} . \tag{C.24}$$

The first difference between directed and undirected networks comes from the fact that we cannot prove this time the inexistence of isolated nodes in the partition of optimal modularity. The previous argumentation was based on the use of (C.12), which now splits in two relationships:

$$\sum_r \tilde{q}_{i,r}^{\text{out}} = -\frac{1}{2w} \left(w_{ii} - \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \right) , \tag{C.25}$$

$$\sum_r \tilde{q}_{j,r}^{\text{in}} = -\frac{1}{2w} \left(w_{jj} - \frac{w_j^{\text{out}} w_j^{\text{in}}}{2w} \right) . \tag{C.26}$$

The next step is the same:

$$\text{if } w_{ii} \leq \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \;\Rightarrow\; \sum_r \tilde{q}_{i,r}^{\text{out}} \geq 0 \;\Rightarrow\; \exists s_1 : \tilde{q}_{i,s_1}^{\text{out}} \geq 0 , \tag{C.27}$$

$$\text{if } w_{ii} \leq \frac{w_i^{\text{out}} w_i^{\text{in}}}{2w} \;\Rightarrow\; \sum_r \tilde{q}_{i,r}^{\text{in}} \geq 0 \;\Rightarrow\; \exists s_2 : \tilde{q}_{i,s_2}^{\text{in}} \geq 0 . \tag{C.28}$$

Since communities s_1 and s_2 need not be the same, the change of modularity (C.24) is not warranted to be positive, and thus isolated nodes are possible in the partition which maximizes modularity. Nevertheless, there exist three kinds of nodes for which we can prove they cannot be isolated in the optimal partition, provided their self-loops are not too large: hairs (nodes that have $s_1 = s_2$) , *sinks* (nodes with only input links) and *sources* (nodes with only output links).

Resum de la Tesi

Introducció

La consolidació d'Internet com a xarxa mundial de comunicació està considerada una de les claus de la revolució tecnològica de finals del segle passat, contribuïnt al fenòmen de la globalització. Les dades confirmen aquest fet; el nombre total d'elements connectats a Internet ha crescut exponencialment any rera any des de 1990, amb més de mil milions d'usuaris utilitzant els seus serveis diàriament. A més, la quantitat de tràfic generat per aquests usuaris també creix entre un 100% i un 1000% anualment, arribant a moure diàriament al voltant de Petabytes (2^{50} bytes) d'informació entre ordinadors de tot el món.

Malgrat que hi ha la idea generalitzada que al darrere del disseny d'Internet hi existeix una gran planificació i un important esforç d'enginyeria, la veritat és que la part tecnològica juga un paper molt petit en el seu desenvolupament; el secret de l'èxit d'aquesta xarxa resideix en un conjunt de protocols i guies tècniques que descriuen com comunicar eficientment tot tipus de dispositius electrònics. A partir d'aquestes guies, entitats independents (com ara els governs o els proveïdors d'Internet) imposen les seves pròpies normatives de com connectar nous dispositius a la xarxa, fomentant la creació d'una infraestructura extremadament heterogènia que ha anat evolucionant durant els últims 20 anys.

Internet presenta dues propietats interessants que en principi no esperaríem d'un sistema amb un creixement tan descentralitzat i sense un disseny preimposat. En primer lloc, és una de les xarxes més robustes que existeixen; tot i el gran nombre d'atacs que pateixen cada dia els seus components, molt pocs atacs han provocat una interrupció de les comunicacions a escala global. En segon lloc, Internet mostra una inusitada eficiència en el lliurament d'informació entre usuaris, tot i l'enorme quantitat de tràfic distribuït pels routers d'arreu del món. Per exemple, des d'un punt de vista estadístic, la latència o els retards soferts pels paquets d'informació són molt baixos en comparació amb altres xarxes tradicionals de comunicació o de transport que tenen un comportament dinàmic similar (com per exemple la xarxa d'autopistes).

Per aquests motius, Internet es considera un dels exemples paradigmàtics del què entenem per sistema complex. Partint del fet que no existeix una definició estandaritzada de complexitat, la comunitat científica sol descriure aquests sistemes utilitzant algunes característiques comuns que podem observar en tots ells, independentment que l'origen dels sistemes sigui social, biològic o tecnològic. Per exemple, a l'igual que passa amb Internet, la majoria de sistemes complexos mostren normalment una organització òptima que apareix sense que hi hagi cap mena de control o disseny extern. Però potser la propietat més destacada que comparteixen els sistemes complexos és la seva no-linealitat, és a dir, que no podem entendre el comportament global del sistema a partir de la suma dels comportaments individuals dels seus components, sinó que existeixen molts altres factors que hem de tenir en compte.

Les xarxes són una de les representacions més utilitzades per a descriure l'estructura de les interaccions entre els elements de qualsevol sistema. Les xarxes més bàsiques es solen modelar utilitzant grafs regulars o aleatoris, els quals mostren un alt grau de similitud quan mesurem qualsevol de les seves parts. Tanmateix, l'observació de les interaccions en sistemes complexos reals demostra que l'estructura de connexions és molt heterogènia, amb una sèrie de característiques no-trivials que no es tenen en compte en les aproximacions anteriors. Aquest grup de xarxes es coneixen per xarxes complexes, és a dir, xarxes que es troben "entre la regularitat i l'aleatorietat", perquè en l'aparent estructura caòtica s'hi troba amagada una organització òptima que facilita el funcionament global dels sistemes complexos.

Amb l'objectiu principal d'entendre i modelar les característiques bàsiques de les xarxes complexes, ha aparescut recentment una nova "'ciència de les xarxes". L'aparició d'aquesta ciència s'ha vist potenciada per tres factors: en primer lloc, l'augment de la potència i la major disponibilitat de recursos computacionals han permès realitzar estudis més detallat i amb xarxes composades per milers (o milions) d'elements. En segon lloc, gràcies a la informatització de les dades i l'aparició d'Internet com un immens repositori d'informació, la comunitat científica pot accedir a una gran quantitat de dades sobre xarxes que cada dia es publiquen en tot tipus de camps. Finalment, el que més ha potenciat l'interès en les xarxes ha estat el desenvolupament d'eines, mesures, i models, ja siguin noves o importades d'altres disciplines, tals com la sociologia, la física estadística o la teoria de grafs.

Paral.lelament a aquests treballs, en els darrers anys algunes de les línies d'investigació de la ciència de les xarxes s'han redirigit cap a l'estudi de quina és la interrelació existent entre l'estructura i els processos dinàmics d'un sistema complex. Els primers estudis han demostrat que existeix una influència

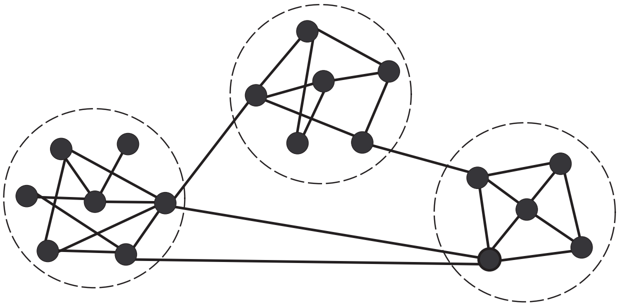


Figure C.1. Exemple de l'estructura de comunitats d'una xarxa. Podem classificar els vertexs en grups on el nombre de links cap a membres del mateix grup es més gran que el nombre de links amb la resta de nodes de la xarxa.

bidireccional entre ambdues parts; per exemple, en el cas particular d'Internet, entendre com l'estructura influencia la dinàmica del flux de dades proporciona una informació molt valuosa sobre com dissenyar millors topologies i protocols de comunicació mes eficients.

Aquesta tesi té com a objectius revisar l'estat de l'art de les tècniques existents per a descriure de les xarxes complexes, i proporcionar noves eines i models que permetin una millor comprensió, tan a nivell topològic com a nivell dinàmic. En particular, hem tractat dos dels problemes que més interès desperten en la literatura actual: el problema de la detecció de l'estructura de comunitats en xarxes complexes (en els dos primer capítols) i el problema la caracterització de les propietats dinàmiques d'un procés de comunicació (en els dos darrers capítols).

La mesoscala de les xarxes complexes

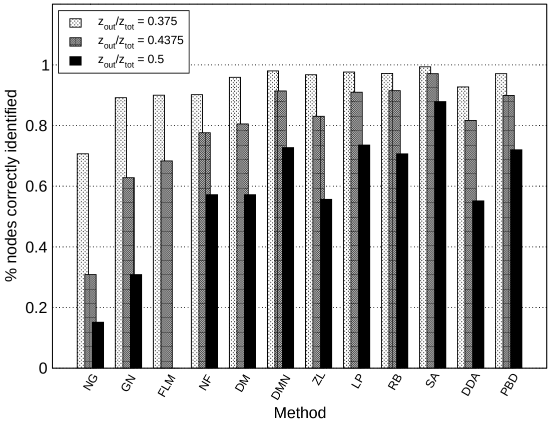
Els nivells de descripció topològica d'una xarxa complexa obtinguts amb les eines que hem comentat permeten la caracterització del nivell microscòpic (a nivell d'un node individual) o bé del macroscòpic (a nivell de tota la xarxa). Al mig d'aquests dos extrems podem, alhora, efectuar una descripció mesoscòpica per a identificar i estudiar les propietats d'aquells grups de nodes que es troben densament connectats, és a dir, grups de nodes on el nombre de connexions amb membres del mateix grup es més gran que el nombre de conexions amb la resta de nodes de la xarxa. El conjunt de tots els grups de nodes que compleixen aquesta condició és el que es coneix com a estructura de comunitats i ens permet oferir una descripció de les escales intermitjes d'una xarxa complexa.

El concepte de comunitat ha estat àmpliament utilitzat en les ciències socials, i reflexa el fet que els individus tendeixen a establir més connexions amb grups de gent amb qui comparteixen aficions, amistat o simplement perquè viuen a prop. No obstant, aquesta organització en comunitats no és exclusiva de les xarxes socials, ja que també es pot observar quan analitzem amb detall diferents sistemes, com ara xarxes metabòliques, la *World Wide Web* o la xarxa mundial d'aeroports (Ravasz et al., 2002, Guimerà et al., 2005). Alguns estudis científics han demostrat que els nodes que pertanyen a la mateixa comunitat acostumen a compartir una sèrie de propietats comuns (Flake et al., 2002, Eckmann and Moses, 2002) i també han demostrat com l'existència d'una estructura de comunitats ben definida influencia els processos dinàmics com ara la sincronització entre els nodes, la difusió de la informació, o l'emergència d'actituts de cooperació entre agents (Arenas et al., 2006b, Arenas et al., 2006a, Lozano et al., 2007).

La identificació i caracterització d'aquestes comunitats de nodes no és una feina trivial. Un dels problemes principals és que la pròpia definició de comunitat està en termes qualitatius, i la determinació de la mesura quantitavia més adecuada es troba encara sota debat. Fins ara, la mesura més acceptada per la comunitat física rep el nom de modularitat (Newman and Girvan, 2004), i mesura quina és la probabilitat de que les conexions internes del grup siguin fruït o no d'un procés aleatòri. L'èxit d'aquesta mesura radica en que proporciona una forma de determinar si una descripció mesoscòpica es més o menys precisa, és a dir, permet mirar dues particions i afirmar quina és millor mirant la que proporciona un valor més alt de modularitat.

Un altre problema que presenta la detecció de comunitats és la gran quantitat de configuracions possibles en que es pot organitzar els nodes. Per a cercar dins d'aquestes configuracions quina és la millor, s'ha desenvolupat un nou conjunt de mètodes capaços de descobrir l'existència de les comunitats a partir de la informació topològica de les xarxes (com es connecten els nodes entre ells). En el capítol dos de la tesi es presenta una extensa comparativa de la literatura actual sobre mètodes de detecció de comunitats, intentat oferir a la comunitat científica un punt de referència en el camp de la detecció. Per a poder comparar aquests mètodes s'han recollit tres conjunts d'eines que ens permeten analitzar la precisió i la velocitat dels mateixos. En primer lloc, utilitzant una sèrie de xarxes generades artificilament (on es controla el nivell de definició de les comunitats), es pot analitzar quin és el nivell de precisió que assoleix un algorisme a l'intentar detectar les comunitats pregenerades. En segon lloc, per a poder decidir la similitud entre dues particions es proposa un mètode basat en la teoria de la informació, conegut com a *mutual information*, que mesura quina és la dependencia mútua existent entre dues configuracions. I en tercer lloc, es proposa un mètode per a comparar la velocitat dels algorismes observant el seu ordre de complexitat, és a dir, com escala el temps d'execució amb el tamany de la xarxa.

A partir dels resultats obtinguts a l'aplicar aquestes eines de comparació als algorismes existents, es pot decidir quin és el mètode que més s'ajusta a la



Label	Order
ZL	$O(n^3)$
LP	$O(n^3)$
NF	$O(n\log^2 n)$
NG	$O(m^2 n)$
GN	$O(n^2 m)$
SA	par. dep.
DM/DMN	$O(n^3)$
WH	$O(n+m)$
PK	$O(\exp(n))$
RB	par. dep.
NS	$O(n^2 \log n)$
DDA	$O(n\log^2 n)$
PBD	$O(n\log^2 n)$

Figure C.2. Dreta: Estudi de la precisió dels algorismes quan els sometem a xarxes amb una estructura de comunitats prefixada. Com més gran es el valor de z_{out}/z_{tot} més difícil resulta trobar les comunitats. Esquerra: Taula on es resumeix com el cost computacional d'alguns dels algorismes descrits en la tesi escala amb el nombre de nodes del sistema n, el nombre de links m i el grau mig $\langle k \rangle$. La correspondència entre etiquetes i algorismes està explicada al capítol 2 de la tesi.

xarxa que vulguem estudiar. El que resulta més sorprenent és que existeix un compromís entre el temps que tarda l'algorisme i la qualitat dels resultats. Per a analitzar xarxa petites i mitjanes (fins a alguns milers de nodes) és recomanable utilitzar els algorismes que proporcionen una detecció més precisa. En canvi, si volem analitzar xarxes més grans, és necessari utilitzar un algorisme més escalable que no podrà garantir que la partició trobada sigui semblant a la més òptima.

Detecció de comunitats utilitzant Extremal Optimization

Al tercer capítol, s'introdueix un mètode alternatiu per a trobar la partició amb la millor modularitat, que intenta superar les limitacions existents en els algorismes descrits anteriorment. Com han demostrat alguns autors, la optimizació de la modularitat és un problema *NP-hard* (Brandes et al., 2007), degut a que l'espai de particions possibles creix més ràpidament que qualsevol potència del tamany del sistema, per la cual cosa l'unica opció disponible per a aproparnos a la partició óptima és utilitzar una cerca heurística que permeti reduir l'espai de les particions possibles a analitzar.

L'algorisme que es proposa en aquest capítol es un mètode divisiu que optimitza la modularitat utilitzant una cerca heurística coneguda per *Extremal Optimization* (EO) (Boettcher and Percus, 2001a, Boettcher and Percus, 2001b). El funcionament bàsic de l'EO consisteix en optimitzar una variable global del sistema (en el nostre cas la modularitat) a partir de millorar la contribució local dels pitjors elements del sistema, mitjançant un procés que implica allaus coevolucionàries. L'eficiència de l'heurística EO s'ha posat de manifest a l'aplicar-se a alguns problemes clàssics (com ara els spin glasses o problemes de coloracio de grafs), millorant els resultats obtinguts per altres heurístiques més consolidades tals com el *simulated annealing* o els algorismes genètics.

Internament, el nostre algorisme EO està dissenyat com una versió més complexa del *Graph Bipartitioning*, un problema clàssic de la teoria de grafs que consisteix en separar una xarxa en dos grups de nodes intentant minimitzar el nombre de links entre els dos grups. Al principi els nodes s'assignen aleatòriament a un dels dos grups, i després es deixa evolucionar el sistema movent els nodes que tenen pitjor contribució a la modularitat total d'un grup a l'altre. En cada pas es mira si el sistema ha assolit una modularitat més alta o no. Un cop es detecta que ens trobem en un punt en que no es pot obtenir una configuració millor, s'eliminen tots els links intermitjos de la xarxa i es torna a començar de nou amb tots els subgrups que hagin quedat. Aquest procés es repeteix recursivament fins que no es pot incrementar més la modularitat de la xarxa.

Després de desenvolupar l'algorisme EO, hem dirigit els esforços en desenvolupar algunes modificacions que ens permeten millorar alguns aspectes puntuals del mateix i alhora poder-lo aplicar en un grup de xarxes mes ampli. En primer lloc, redefinint la formulació de modularitat i fent uns canvis menors en el codi hem creat un dels primers mètodes capaç de poder analitzar xarxes dirigides i pesades. En segon lloc, hem aplicat petits canvis a nivell algorísmic per a solucionar alguns problemes relacionats amb la recursivitat, permetent al sistema assolir valors més alts de modularitat. I en tercer lloc, hem proposat un mètode que permet reduir el tamany de la xarxa preservant la modularitat de la millor partició. Aquesta reducció permet que qualsevol algorisme basat en optimitzar la modularitat pugui analitzar amb més detall l'espai de configuracions possibles, i per tant poder obtenir millors configuracions utilizant un menor temps d'anàlisi.

Els resultats presentats al final del capítol mostren que el nostre algorisme esdevé una de les millors alternatives per a identificar l'estructura de comunitats d'una xarxa complexa. Els valors de la modularitat obtinguts a l'analitzar les principals xarxes de referència se situen entre els més alts publicats en la literatura sobre comunitats. Per altra banda, tot i no ser un dels algorismes més ràpids, el temps d'anàlisi escala com $O(n^2 log(n))$ amb el tamany de la xarxa, permetent realitzar la detecció de comunitats de forma acurada en xarxes mitjanes i grans.

Finalment, s'ha aplicat l'algorisme per a estudiar una xarxa real, la xarxa de projectes FP6 de la comunitat eruopea. L'anàlisi detallat de les comunitats trobades demostra que identifiquem clarament grups de companyies i institucions amb un perfil similar com, per exemple, empreses relacionades amb el sector automobilístic.

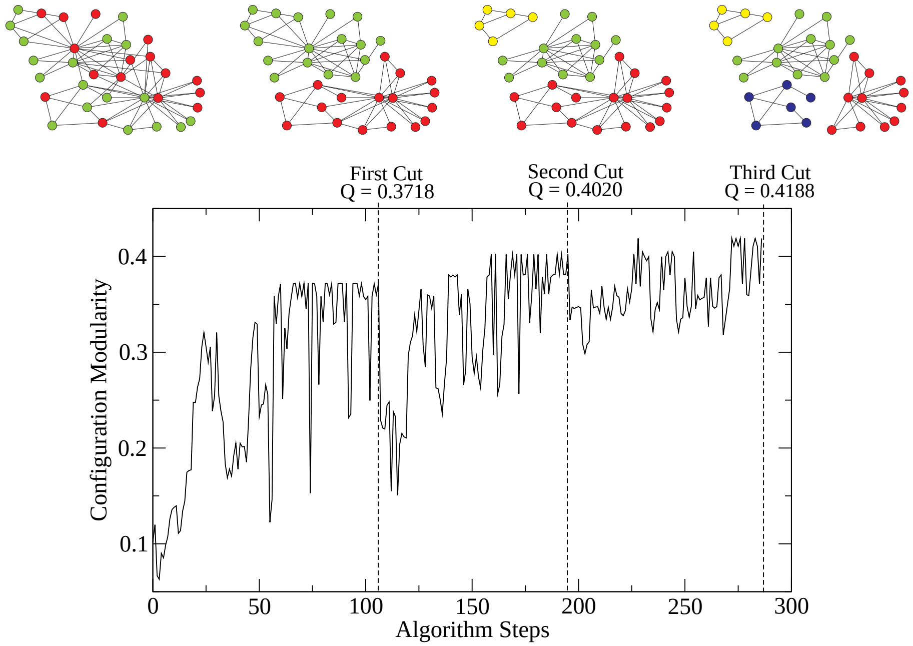


Figure C.3. Exemple del funcionament del nostre algorisme de detecció de comunitats quan analitzem la xarxa Zachary, una de les més utilitzades per a comprovar la precisió de la majoria de mètodes. Gràfic superior: Estat dels nodes de la xarxa despres de la inicialització aleatoria en dos grups i després en cadascun dels moments en que l'algorisme va tallant la xarxa recursivament. Gràfic inferior: Evolució del valor de la modularitat en cadascun dels passos del procés evolutiu. Les barres de separació signifiquen que hem arribat a un estat estacionari i per tant procedim a subdividir el graf en els talls que observem a la part superior.

Estudi de les fluctuacions del tràfic en una xarxa complexa

Recentment, els estudis sobre xarxes complexes han començat a estudiar les propietats dels processos dinàmics que tenen lloc sobre aquestes xarxes. En el nostre cas ens hem centrat únicament en els processos de comunicació, amb l'intenció d'entendre els paràmetres que governen el fluxe de paquets que es mouen utilitzant la xarxa complexa, esbrinant quina és la relació entre l'estructura de la xarxa i el comportament d'aquests paquets.

Els principals resultats obtinguts fins ara al voltant de l'estudi del fluxe de tràfic es referèixen a quines son les causes que introdueixen la congestió en el sistema. No obstant, l'observació del comportament del tràfic en algunes xarxes reals (com per exemple Internet) mostra que el tràfic no està governat pels processos de congestió, sinó que és un tràfic amb un comportament normal i que es troba sotmès a grans fluctuacions que el poden portar a congestionar en moments puntuals. Això ha obert la porta a un nou grup d'estudis que han intentat caracteritzar les fluctuacions del tràfic en varis sistemes complexos, com ara la xarxa d'autopistes, la xarxa fluvial o la mateixa Internet (de Menezes and Barabási, 2004a, de Menezes and Barabási, 2004b). Tots aquests sistemes es poden representar a un nivell abstracte com xarxes on una serie de packets viatgen entre els seus nodes. En particular, els autors relacionen quin és el tràfic mig $\langle f \rangle$ en cadascun dels nodes amb la seva variabilitat σ, i descobreixen que existeix una relació d'escala entre els dos valors,$\langle f \rangle^{\alpha} \sim \sigma$. A més, proposen que l'exponent α és capaç de caracteritzar les fluctuacions del sistema, i que aquest valor pot ser únicament $1/2$ o be 1.

El principal problema d'aquests treballs és que els autors no tenen en compte la possibilitat que els paquets interactuin entre sí, evitant justament l'aparició de la congestió en el sistema. Per a entendre millor les fluctuacions, en el capítol quatre de la tesi es proposa un nou model per a estudiar aquestes fluctuacions i per a comprovar si existeixen unicament dos possibles valors per a l'exponent d'escala. El model està basat en un procés dinàmic de comunicació on cada node té una capacitat limitada per a enviar i rebre paquets. Quan el node està ocupat, els paquets s'esperen en una cua fins a poder ser servits. Per a simplificar l'estudi s'utilitzen cues del tipus M/M/1 que es troben governades per distribucions Poisson (Allen, 1990).

L'anàlisi dels resultats obtinguts en el nostre model mostra que modificant alguns dels paràmetres podem provocar una transició de l'exponent entre $1/2$ i 1. En primer lloc, es pot comprovar que si la mida de la finestra de mostreig és més petit que el temps mitjà que transcorre entre que un node rep dos paquets consecutius, l'exponent α sempre serà $1/2$, independentment de les fluctuacions reals del tràfic. En canvi, si la mida de la finestra és suficientment gran,

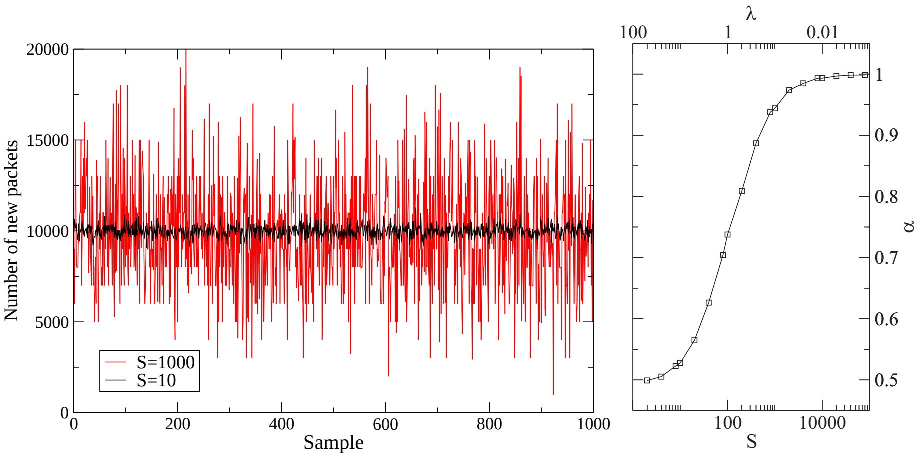


Figure C.4. Dreta: Exemple de dos sistemes que tenen el mateix tràfic mitjà però diferent variabiliat. Esquerra: Transició que es produeix en l'exponent d'escala α a mesura que anem introduint més variabilitat en el sistema.

el valor de l'exponent serà funció de les fluctuacions reals del sistema. En segon lloc, s'observa que si mantenim el nombre de paquets que hi ha en el sistema en un moment donat, però es canvia el temps que esta un paquet actiu i la ràtio de creació de nous paquets, es pot introduir un altre cop una transició entre els dos exponents. Finalment, quan s'afegeix la possibilitat de que existeixi congestió en el sistema, torna a aparèixer una transició entre $1/2$ i 1. Quan ens apropem al punt crític de la congestió observem una transició de fase cap a $\alpha = 1/2$. L'explicació és que un cop el sistema es troba congestionat, el valor de l'exponent passa a ser únicament funció de la variança de la distribució Poisson.

Per tant, utilitzant aquest model es pot afirmar que els dos exponents universals predits pel treball de Menezes i Barabasi no es corresponen amb el que el nostre model prediu. Per a corroborar aquesta afirmació hem realitzat l'estudi de les fluctuacions d'una xarxa real, la xarxa Abilene que composa el nucli del que es coneix com a Internet 2. Al caracteritzar les fluctuacions dels tràfic en els 112 nodes de la xarxa s'observen exponents que varien entre 0.71 i 0.86, demostrant que els sistemes reals poden tenir exponents diferents de $1/2$ i 1.

Robustesa dinàmica d'un proces de comunicacio

En el darrer capítol de la tesi hem centrat l'atenció en una altra propietat molt interessant de les xarxes complexes: la seva robustesa davant la fallida d'alguns dels seus components. La robustesa d'un sistema és un element clau per a mantenir el funcionament dels processos dinàmics que hi tenen lloc. Per exemple, en el cas d'Internet, l'estabilitat dels sistema és un factor clau per a garantir la màxima eficiència de la xarxa, és a dir, poder mantenir el temps mitjà que es triga en enviar la informació i evitar la pèrdua de paquets de dades.

Els estudis tradicionals han analitzat quins són els efectes que comporta eliminar alguns dels components de la xarxa, ja sigui de forma intencionada o aleatòria, en les propietats estructurals de la xarxa. La majoria d'aquests estudis defineixen la robustesa d'una xarxa com la capacitat de mantenir una component connexa del mateix ordre que el tamany del sistema. No obstant, en les xarxes complexes es pot donar el cas que tinguem una xarxa que estigui connectada però el funcionament dels processos dinàmics hagi canviat a causa de que s'hagin eliminat alguns components clau. En el cas concret d'una xarxa de comunicació, podria ser que els nodes estiguin connectats, però que al treure nodes apareixi congestió en el sistema i, per tant, que el rendiment del sistema disminueixi.

En aquest escenari, en el capítol 5 hem introduït el concepte de robustesa dinàmica d'una xarxa complexa, definida com la capacitat de mantenir el sistema funcionant quan alguns dels nodes fallen. Per estudiar la robustesa dinàmica d'un procés de comunicació hem utilitzat un model de tràfic semblant al presentat en el capítol previ. A partir d'aquest model hem analitzat l'efecte que provoca l'eliminació aleatòria de nodes de la xarxa en la capacitat màxima del sistema per a distribuïr la informació, mesurant els canvis que sofreix el punt crític de la congestió. Hem analitzat què passa quan realitzem l'experiment en diferents tipus de xarxes (xarxes aleatòries i xarxes amb una distribució de grau scale-free) i utilizant un protocol d'enrutament amb diferents graus de coneixement de la xarxa, des d'un protocol aleatori (amb coneixement zero) a un protocol basat en camins mínims (amb coneixement total).

Els resultats de l'estudi mostren que en les xarxes scale-free l'eliminació de nodes sempre augmenta la capacitat màxima del sistema, en les xarxes regulars la capacitat decreix considerablement i finalment en les xarxes aleatòries el canvi en la capacitat es troba en funció del radi de coneixement del protocol. També hem realitzat una aproximació teòrica utilitzant la descripció del punt crític de la congestió presentada en (Guimerà et al., 2002b). Tant l'anàlisi experimental com l'aproximació teòrica ens han permès determinar quins són els principals paràmetres que controlen aquests canvis en la congestió: la centralitat del node més important (aquell pel qual hi passa la major quantitat de paquets) i sobretot l'efecte que suposa l'eliminació de nodes en la distància mitjana que recorre un paquet per la xarxa.

Conclusions

L'ús de les xarxes complexes per a representar les interaccions d'un sistema complex ha estat una peça clau per a poder treballar amb el gran nombre de sistemes biològics, tecnològics, econòmics o socials que contínuament es publiquen. Els avenços en el que coneixem per "ciència de les xarxes" han estat encaminats des d'un bon principi a contestar dues grans preguntes: què podem aprendre d'un sistema si ens mirem la seva estructura interna i quina és la relació que existeix entre una determinada estructura i el comportament dinàmic del sistema suportat. En aquesta tesi hem treballat en alguns dels problemes principals d'ambdues qüestions, intentant proporcionar un coneixement més profund de les xarxes complexes que permetrà a la comunitat científica entendre millor quin és el funcionament d'un sistema complex.

Durant els darrers anys s'ha posat de manifest la importància de la detecció de comunitats com un element clau per a caracteritzar l'organització interna d'una xarxa. La identificació d'aquestes escales intermedies ha permès a la comunitat científica entendre com els elements d'un sistema s'agrupen per a formar comunitats funcionals i alhora analitzar la influència d'aquestes comunitats en el comportament global.

El problema de la detecció de comunitats ha atret l'atenció de científics provinents de camps molt diversos, com es demostra en el gran nombre de treballs per intentar detectar i quantificar aquesta estructura que s'han publicat en els darrers cinc anys. Tot i que els mètodes que hem analitzat en el segon capítol fan servir aproximacions completament diferents, la precisió obtinguda quan detecten una estructura de comunitats ben definida és bastant similar. Per aquesta raó es pot afirmar que el problema de la detecció de comunitats és un exemple paradigmàtic d'un problema que es pot enfocar des de molts punts de vista diferents i, a la vegada, arribar a les mateixes conclusions. Si s'augmenta la dificultat de trobar les comunitats, s'observa l'existència d'una limitació entre el temps d'execució i la precisió dels algorismes. Aquest és un problema comú que apareix cada cop que volem treballar amb quantitats de dades molt grans. En el cas de la detecció de comunitats, els mètodes més precisos normalment no poden treballar amb xarxes mes grans de 10000 nodes. Si per altra banda es vol analitzar una xarxa més gran, es necessita utilitzar una de les alternatives més ràpides perdent, aleshores, la precisió que ens garanteixen els mètodes més acurats.

Al tercer capítol hem presentat un mètode de detecció de comunitats basat en l'heurística extremal optimization. Tot i que aquesta heurística no és tan popular com d'altres (p. ex. simulated annealing), s'ha demostrat que dóna molts bons resultats quan l'apliquem a problemes clàssics d'optimització. Addicionalment, entès que darrera d'aquesta heurística hi existex un procés evolucionari on els nodes s'autoagrupen entre ells fins que arriben a un estat estacionari, l'extremal optimization es pot fer servir, alhora, per a entendre els processos que existeixen darrere la creació de les comunitats. Al comparar els resultats del nostre mètode amb els analitzats en el segon capítol, es pot observar que el nostre es troba entre els més precísos, amb l'avantatge que realitza la cerca en un temps menor. Per aquestes raons, els resultats obtinguts amb el nostre algorisme han estat un punt de referència a l'hora de comparar els mètodes de detecció publicats a posteriori.

En aquest capítol també hem realitzat una anàlisi més profunda de la formulació de la modularitat. Hem redefinit la mesura per a poder tractar xarxes dirigides i pesades i hem estudiat quina es la contribució que tenen algunes subestructures concretes al valor global de la modularitat. A partir d'aquest coneixement, hem proposat algunes millores per als mètodes de detecció. Per una banda, hem proposat un seguit de modificacions algorismiques que permeten optimitzar la precisió dels algorismes recursius fins a un 20%. Per altra banda, hem proposat un mètode que ens permet reduïr la mida d'una xarxa fins a un 40% sense alterar la modularitat de les configuracions, permeten un anàlisi molt més ràpid i precís.

En els capítols quatre i cinc hem estudiat algunes propietats dinàmiques dels processos de comunicació basats en xarxes complexes. Creiem que es necessari obrir i reforçar les línies de recerca dedicades a l'estudi de les propietats globals dels processos dinàmics. Quan els científics volen modelar la topologia d'una xarxa complexa particular (per exemple Internet), fan servir les eines descrites als primers capítols per a mesurar quines són les propietats estructurals que presenta la xarxa (p.ex. distribucio de grau scale-free) i llavors reprodueixen aquestes propietats en els models obtenint un model simple però acurat de la realitat. En canvi, quan es vol modelar un proces dinàmic no poden seguir aquestes mateixes passes, ja que manca un conjunt d'eines equivalent al que disposem per a la topologia.

En aquest context, al quart capítol hem estudiat les fluctuacions del tràfic en una xarxa complexa per a proporcionar una caracterització global del seu flux. A l'article de Menezes i Barabasi, els autors proposen l'existència de dues classes universals que permeten caracteritzar la relació entre mitja i desviació típica amb uns exponents $\alpha = 1/2$ and $\alpha = 1$. Tot i que creiem que aquests dos valors són importants, ja que sembla que actuen com a límits dels valors observats en sistemes reals, pensem que aquestes valors no són únics. L'anàlisi d'un model de tràfic molt simple basat en un sistema de cues Poisson revela tres mecanismes que provoquen una transició entre els dos exponents. Per a corroborar l'existència d'exponents intermedis en sistemes reals hem estudiat les fluctuacions que hi ha en el tràfic d'Internet de la xarxa Abilene. L'estudi mostra que el tràfic d'aquesta xarxa es pot caracteritzar a partir d'un exponent d'escala $\alpha \sim 0.75$. Aquests resultats estan en concordància amb altres estudis de les fluctuacions dels sistemes complexos presentats en el review d'Eisler *et al.*, on els autors troben una gran varietat d'exponents en diferents sistemes complexos.

La capacitat de mantenir la comunicació entre dos nodes de la xarxa quan alguns d'ells fallen és una altra de les caracteristiques rellevants de les xarxes complexes. A diferència de la majoria d'estudis realitzats, creiem que és més important estudiar la incidència de les fallides dels nodes en els processos dinàmics suportats per la xarxa que centrar-se en l'efecte sobre en la topologia. Al capítol cinquè hem definit la robustesa dinàmica d'un procés de comunicació com la capacitat del tràfic per a evitar la congestió quan eliminem una fracció dels nodes de la xarxa. Fent servir un model de tràfic semblant a l'utilitzat en el capítol quatre, hem proporcionat un nou punt de vista de com es comporta el punt crític de la congestió quan eliminem nodes aleatòriament, analitzant diferents topologies i protocols d'enrutament. Finalment, a partir d'un estudi teòric del problema hem analitzat quines són les causes darrere els canvis en la capacitat màxima.

Publication list

Community Structure in complex networks

- Duch, J., and Arenas, A. (2005), Community Detection in complex networks using extremal optimization, *Phys. Rev. E* 72, 027104. Also in Virtual Journal of Biological Physics Research, September 2005.

- Danon, L., Duch, J., Díaz-Guilera, A., and Arenas, A. (2005), Comparing community structure identification, *JSTAT*, P09008.

- Danon, L., Duch, J., Díaz-Guilera, A., and Arenas, A., Community structure identification in "Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science", World Scientific, June 2007

- Lozano, S., Duch, J., and Arenas, A., (2006) Community detection in a large social dataset of European projects, *Workshop on Link Analysis, Counterterrorism and Security (SIAM on Data mining 2006)*, Washington USA, 2006.

- Lozano, S., Duch, J., and Arenas, A., (2007) Analysis of large social datasets by community detection, *European Physical Journal ST*, vol. 143, 257-259

- Arenas, A., Duch, J., Fernández, A., and Gómez, S., (2007) Size reduction of complex networks preserving modularity, *New Journal of Physics*, Vol. 9, 176.

Traffic in complex networks

- Duch, J., and Arenas, A. (2006), Scaling of Fluctuations in Traffic on Complex Networks, *Phys. Rev. Lett.* 96, 218702.

- Duch, J., and Arenas, A. (2007), A model to study the scaling of traffic fluctuations on complex networks, *European Physical Journal ST*, vol. 143, 253-255

- Duch, J., and Arenas, A. (2007), Effect of random failures on traffic in complex networks, *Proceedings of SPIE*, Volume 6601, 66010O.

- Duch, J., Díaz-Guilera, A., and Arenas, A. (2007), Congestion in traffic on complex networks under random failures, in preparation.

References

[Aho et al., 1983] Aho, D. V., Ullman, J. D., and Hopcroft, J. E. (1983). *Data Structures and Algorithms*. Addison-Wesley.

[Albert et al., 2000] Albert, R., Jeong, H., and Barabási, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:376.

[Allen, 1990] Allen, O. (1990). *Probability, Statistics and Queueing Theory with Computer Science Application*. Academic Press.

[Alvarez-Hamelin et al., 2007] Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., and Vespignani, A. (2007). k-core decomposition: a tool for the analysis of large scale internet graphs. *arXiv.org:cs/0511007*.

[Amaral and Ottino, 2004] Amaral, L. A. N. and Ottino, J. (2004). Complex networks: Augmenting the framework for the study of complex systems. *Eur. Phys. J. B*, 38:147–162.

[Andersen et al., 2002] Andersen, D. G., Feamster, N., Bauer, S., and Balakrishnan, H. (2002). Topology inference from bgp routing dynamics. In *2nd Internet Measurement Workshop*.

[Arenas et al., 2004] Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P. M., and Guimerà, R. (2004). Community analysis in social networks. *Eur. Phys. J. B*, 38:373–380.

[Arenas and Díaz-Guilera, 2007] Arenas, A. and Díaz-Guilera, A. (2007). Synchronization and modularity in complex networks. *Eur. Phys. J. B*, 143:19–25.

[Arenas et al., 2001] Arenas, A., Díaz-Guilera, A., and Guimerà, R. (2001). Communication in networks with hierarchical branching. *Phys. Rev. Lett.*, 86(14):3196–3199.

[Arenas et al., 2006a] Arenas, A., Díaz-Guilera, A., and Pérez-Vicente, C. (2006a). Synchronization processes in complex networks. *Physica D*, 224:27–34.

[Arenas et al., 2006b] Arenas, A., Díaz-Guilera, A., and Pérez-Vicente, C. J. (2006b). Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.*, 96:114102.

[Arenas et al., 2007] Arenas, A., Fernández, A., and Gómez, S. (2007). Multiple resolution of the modular structure of complex networks. *physics/0703218v1*.

[Bagrow and Bollt, 2005] Bagrow, J. P. and Bollt, E. M. (2005). Local method for detecting communities. *Phys. Rev. E*, 72:046108.

[Bak, 1996] Bak, P. (1996). *How Nature Works*. Copernicus.

[Bak and Sneppen, 1993] Bak, P. and Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, 71:4083–4086.

[Bak et al., 1987] Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: an explanation of 1 / f noise. *Phys. Rev. Lett.*, 59:381–384.

[Banavar et al., 1987] Banavar, J. R., Sherrington, D., and Sourlas, N. (1987). Graph bipartitioning and statistical mechanics. *J. Phys. A: Math. Gen.*, 20:L1–L8.

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergenge of scaling in random networs. *Science*, 286:509–512.

[Barabási and Albert, 2002] Barabási, A. L. and Albert, R. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

[Barrat et al., 2004a] Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004a). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci*, 101:3747–3752.

[Barrat et al., 2004b] Barrat, A., Barthélemy, M., and Vespignani, A. (2004b). Modeling the evolution of weighted networks. *Phys. Rev. E*, 70:066149.

[Barthelemy, 2004] Barthelemy, M. (2004). Betweenness centrality in large complex networks. *Eur. Phys. J. B*, 38:163.

[Barthelemy and Flammini, 2006] Barthelemy, M. and Flammini, A. (2006). Optimal traffic networks. *J. Stat. Mech.: Theor. Exp.*, 07:L07002.

[Bender and Canfield, 1978] Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theory A*, 24:296–307.

[Blatt et al., 1996] Blatt, M., Wiseman, S., and Domany, E. (1996). Superparamagnetic clustering of data. *Phys. Rev. Lett.*, 76:3251–3254.

[Boccaletti et al., 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308.

[Boettcher and Percus, 2000] Boettcher, S. and Percus, A. G. (2000). Nature's way of optimizing. *Artificial Intelligence*, 119(1-2):275–286.

[Boettcher and Percus, 2001a] Boettcher, S. and Percus, A. G. (2001a). Extremal optimization for graph partitioning. *Phys. Rev. E*, 64:026114.

[Boettcher and Percus, 2001b] Boettcher, S. and Percus, A. G. (2001b). Optimization with extremal dynamics. *Phys. Rev. Lett.*, 86(23):5211–5214.

[Boettcher and Percus, 2002] Boettcher, S. and Percus, A. G. (2002). Extremal optimization: an evolutionary local-search agorithm. *cs/0209030v1*.

[Boguñá et al., 2004] Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., and Arenas, A. (2004). Models of social networks based on social distance attachment. *Phys. Rev. E*, 70:056122.

[Bollobas, 1983] Bollobas, B. (1983). The evolution of sparse graphs. In *Graph Theory and Combinatorics, conference in honor of Paul Erdos*.

[Bollobas, 1998] Bollobas, B. (1998). *Modern Graph Theory*. Springer, New York.

[Boss et al., 2004] Boss, M., Elsinger, H., Summer, M., and Thurner, S. (2004). The network topology of the interbank market. *Financial Stability Report*, 7.

[Brandes et al., 2007] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On finding graph clusterings with maximum modularity. In *Proc. 33rd Intl. Workshop Graph-Theoretic Concepts in Computer Science*, volume 4769, pages 121–132.

[Bron and Kerbosch, 1973] Bron, C. and Kerbosch, J. (1973). Finding all cliques in an undirected graph. *Communications of the ACM*, pages 575–577.

[Bu and Towsley, 2002] Bu, T. and Towsley, D. (2002). On distinguishing between internet power law topology generators. In *Proceedings of INFOCOM 2002*.

[Callaway et al., 2000] Callaway, D. S., Newman, M. E. J., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: percolation in random graphs. *Phys. Rev. Lett.*, 85:5468–5471.

[Capocci et al., 2004] Capocci, A., Domenico, V., Servedio, P., Caldarelli, G., and Colaiori, F. (2004). Communities detection in large networks. In *Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, Rome, Italy, October 16, 2004, Proceeedings*, volume 3243 of *Lecture Notes in Computer Science*, pages 181–188.

[Carmi et al., 2007] Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., and Shir, E. (2007). A model of internet topology using k-shell decomposition. *Proc. Natl. Acad. Sci*, 104:11150–11154.

[Chang et al., 2001] Chang, H., Jamin, S., and Willinger, W. (2001). Inferring as-level internet topology from router-level path traces. *Proc. SPIE*, 4526:196–207.

[Chang et al., 2006] Chang, H., Roughan, M., Uhlig, S., Alderson, D., and Willinger, W. (2006). The many facets of internet topology and traffic. *Networks and Heterogeneus Media*, 1:596–600.

[Clauset and Moore, 2005] Clauset, A. and Moore, C. (2005). Accuracy and scaling phenomena in internet mapping. *Phys. Rev. Lett.*, 94:018701.

[Clauset et al., 2004] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

[Cohen et al., 2002] Cohen, R., ben Avraham, D., and Havlin, S. (2002). Percolation critical exponents in scale-free networks. *Phys. Rev. E*, 66:036113.

[Cohen et al., 2000] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4629.

[Cohen et al., 2001] Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. (2001). Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685.

[Cohen and Havlin, 2003] Cohen, R. and Havlin, S. (2003). Scale-free networks are ultrasmall. *Phys. Rev. Lett.*, 90:058701.

[Cohen and Raz, 2006] Cohen, R. and Raz, D. (2006). The internet dark matter - on the missing links in the as connectivity map. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications*.

[Cormen et al., 1990] Cormen, T., Leiserson, C., Rivest, R., and Stein, C. (1990). *Introduction to Algorithms, Second Edition*. MIT Press and McGraw-Hill.

[Crucitti et al., 2003] Crucitti, P., Latora, V., Marchiori, M., and Rapisarda, A. (2003). Efficiency of scale-free networks: error and attack tolerance. *Physica A*, 320:622–642.

[Crucitti et al., 2004] Crucitti, P., V.Latora, and Marchiori, M. (2004). Model for cascading failures in complex networks. *Phys. Rev. E*, 69:045104.

[da Fontoura Costa et al., 2007] da Fontoura Costa, L., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physica*, 56:167–242.

[Dall'Asta et al., 2005] Dall'Asta, L., Alvarez-Hamelin, I., Barrat, A., Vázquez, A., and Vespignani, A. (2005). Statistical theory of internet exploration. *Phys. Rev. E*, 71:036135.

[Danon et al., 2006] Danon, L., Díaz-Guilera, A., and Arenas, A. (2006). The effect of size heterogeneity on community identification in complex networks. *J. Stat. Mech.: Theor. Exp.*, page P11010.

[de Menezes and Barabási, 2004a] de Menezes, M. A. A. and Barabási, A. L. (2004a). Fluctuations in network dynamics. *Phys. Rev. Lett*, 92:028701.

[de Menezes and Barabási, 2004b] de Menezes, M. A. A. and Barabási, A. L. (2004b). Separating internal and external dynamics of complex systems. *Phys. Rev. Lett.*, 93:068701.

[Dimitropoulos et al., 2007] Dimitropoulos, X., Krioukov, D., Fomenkov, M., Huffaker, B., Hyun, Y., k. claffy, and Riley, G. (2007). As relationships: Inference and validation. *ACM SIGCOMM Computer Communication Review (CCR)*, 37:29–40.

[Doar, 1996] Doar, M. B. (1996). A better model for generating test networks. In *Proceedings of the IEEE Global Telecommunications Conference*.

[Dodds et al., 2003] Dodds, P. S., Muhamad, R., and Watts, D. (2003). An experiment study of search in global social networks. *Science*, 301:827–829.

[Donetti and Muñoz, 2004] Donetti, L. and Muñoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *J. Stat. Mech.: Theor. Exp.*, page P10012.

[Donetti and Muñoz, 2005] Donetti, L. and Muñoz, M. A. (2005). Improved spectral algorithm for the detection of network communities. In *AIP Conference Proceedings*, volume 779, pages 104–107.

[Donnet et al., 2005] Donnet, B., Friedman, T., and Crovella, M. (2005). Improved algorithms for network topology discovery. *Lecture Notes in Computer Science*, 3431:149–162.

[Dorogovtsev and Mendes, 2002] Dorogovtsev, S. and Mendes, J. F. F. (2002). Evolution of networks. *Adv. Phys.*, 51:1079–1187.

[Doyle et al., 2005] Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R., and Willinger, W. (2005). The "robust yet fragile" nature of the internet. *Proc. Natl. Acad. Sci*, 102:14497–14502.

[Echenique et al., 2004] Echenique, P., Gómez-Gardeñes, J., and Moreno, Y. (2004). Improved routing strategies for internet traffic delivery. *Phys. Rev. E*, 70:056105.

[Eckmann and Moses, 2002] Eckmann, J.-P. and Moses, E. (2002). Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proc. Natl. Acad. Sci.*, 99(9):5825–5829.

[Eisler et al., 2007] Eisler, Z., Bartos, I., and Kertesz, J. (2007). Fluctuation scaling in complex systems: Taylor's law and beyond. *arXiv:0708.2053*.

[Eisler et al., 2005] Eisler, Z., Kertesz, J., Yook, S.-H., and Barabási, A.-L. (2005). Multiscaling and non-universality in fluctuations of driven complex systems. *Europhys. Lett.*, 69:664–670.

[Eldredge and Gould, 1972] Eldredge, N. and Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In Schopf, T., editor, *Models in Paleobiology*, pages 82–11. San Francisco: Freeman Cooper.

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathemticae (Debrecen)*, 6:290–297.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. *Computer Communications Review*, 29:251–262.

[Farkas et al., 2007] Farkas, I., Ábel, D., Palla, G., and Vicsek, T. (2007). Weighted network modules. *New Journal of Physics*, 9:180.

[Fiedler, 1973] Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(3):298–305.

[Flake et al., 2002] Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of communities. *IEEE Computer*.

[Fortunato and Barthélemy, 2007] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci.*, 104:36–41.

[Fortunato et al., 2004] Fortunato, S., Latora, V., and Marchiori, M. (2004). Method to find community structures based on information centrality. *Phys. Rev. E*, 70:056104.

[Fred and Jain, 2003] Fred, A. L. N. and Jain, A. K. (2003). Robust data clustering. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, USA*, pages II–128–133.

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41.

[Fukuda et al., 1999] Fukuda, K., Takayasu, H., and Takayasu, M. (1999). Observation of phase transition phenomen in internet traffic. *Advances in Perfomance Analysis*, 2:21–44.

[Fukuda et al., 2000] Fukuda, K., Takayasu, H., and Takayasu, M. (2000). Origin of critical behavior in ethernet traffic. *Physica A*, 297:289–301.

[Gallos et al., 2005] Gallos, L., Cohen, R., Argyrakis, P., Bunde, A., and Havlin, S. (2005). Stability and topology of scale-free networks under attack and defense strategies. *Phys. Rev. Lett.*, 94:188701.

[Gao, 2000] Gao, L. (2000). On inferring autonomous system relationships in the internet. In *Proc. IEEE Global Internet Symposium,2002*.

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99(12):7821–7826.

[Gleiser and Danon, 2003] Gleiser, P. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573.

[Glover, 1986] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13:533 – 549.

[Goh et al., 2001] Goh, K., Kahng, B., and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Phys Rev Lett.*, 87:278701.

[Goh et al., 2005] Goh, K., Noh, J., Kahng, B., and Kim, D. (2005). Load distribution in weighted complex networks. *Phys. Rev. E*, 72:017102.

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional.

[Golub and van Loan, 1996] Golub, G. H. and van Loan, C. F. (1996). *Matrix Computation*. Johns Hopkins University Press.

[Gordon et al., 2007] Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Richardson, R. (2007). Csi/fbi computer crime and security survey. Technical report, Computer Security Institute.

[Govindan and Radoslavov, 2002] Govindan, R. and Radoslavov, P. (2002). An analysis of the internal structure of large autonomous systems. Technical report, Technical Report 02-777, Computer Science Department, University of Southern California.

[Guardiola et al., 2002] Guardiola, X., Guimerà, R., Arenas, A., Díaz-Guilera, A., and Amaral, L. A. N. (2002). Micro- and macro-structure of trust networks. *cond-mat/0206240*.

[Guimerà and Amaral, 2005a] Guimerà, R. and Amaral, L. A. N. (2005a). Cartography of complex networks: modules and universal roles. *J. Stat. Mech.: Theor. Exp.*, P02001.

[Guimerà and Amaral, 2005b] Guimerà, R. and Amaral, L. A. N. (2005b). Functional cartography of complex metabolic networks. *Nature*, 433:895–900.

[Guimerà et al., 2002a] Guimerà, R., Arenas, A., Díaz-Guilera, A., and Giralt, F. (2002a). Dynamical properties of model communication networks. *Phys. Rev. E*, 66:026704.

[Guimerà et al., 2003] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., and Arenas, A. (2003). Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103(R).

[Guimerà et al., 2002b] Guimerà, R., Díaz-Guilera, A., Vega-Redondo, F., Cabrales, A., and Arenas, A. (2002b). Optimal network topologies for local search with congestion. *Phys. Rev. Lett.*, 89(24):248701.

[Guimerà et al., 2005] Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci.*, 102:7794–7799.

[Guimerà et al., 2004] Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70(025101).

[Hawkinson and Bates, 1996] Hawkinson, J. and Bates, T. (1996). Rfc1930: Guidelines for creation, selection, and registration of an autonomous system (as). Technical report, IETF.

[Holland, 1975] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

[Holme et al., 2002a] Holme, P., Huss, M., and Jeong, H. (2002a). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532Ű538.

[Holme et al., 2002b] Holme, P., Kim, B., Yoon, C., and Han, S. (2002b). Attack vulnerability of complex networks. *Phys. Rev. E*, 65:056109.

[Holme and Kim, 2002] Holme, P. and Kim, B. J. (2002). Vertex overload breakdown in evolving networks. *Phys. Rev. E*, 65:066109.

[Hopcroft et al., 2004] Hopcroft, J., Khan, O., Kulis, B., and Selman, B. (2004). Tracking evolving communities in large linked networks. *Proc. Natl. Acad. Sci.*, 101:5249–5253.

[Huberman and Lukose, 1997] Huberman, B. A. and Lukose, R. M. (1997). Social dilemmas and internet congestion. *Science*, 25:535.

[Huffaker et al., 2000] Huffaker, B., Fomenkov, M., Moore, D., Nemeth, E., and Claffy, K. (2000). Measurements of the internet topology in the asia-pacific region. In *INET'00, Yokohama, Japan, The Internet Society*.

[Huffaker et al., 1998] Huffaker, B., Plummer, D., Moore, D., and k. claffy (1998). Topology discovery by active probing. Technical report, Cooperative Association for Internet Data Analysis - CAIDA.

[Ijiri and Simon, 1977] Ijiri, Y. and Simon, H. A. (1977). *Skew distributions and the sizes of business firms*. North-Holland.

[Jackson, 1957] Jackson, J. (1957). Networks of waiting lines. *Oper. Res.*, 5:518–251.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407:651–654.

[Karagiannis et al., 2004] Karagiannis, T., Molle, M., Faloutsos, M., and Broido, A. (2004). A nonstationary poisson view of internet traffic. In *IEEE Conference on Computer Communications (INFOCOM)*.

[Kernighan and Lin, 1970] Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell System Tech J*, 49:291–307.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gilatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220.

[Knuth, 1993] Knuth, D. E. (1993). *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA.

[Kumpula et al., 2007] Kumpula, J. M., Saramaki, J., Kaski, K., and Kertész, J. (2007). Limited resolution in complex network community detection with potts model approach. *cond-mat/0610370v2*.

[Kuncheva and Hadjitodorov, 2004] Kuncheva, L. I. and Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Systems, Man and Cybernetics, 2004 IEEE International Conference*, volume 2, pages 1214–1219.

[Lakhina et al., 2002] Lakhina, A., Byers, J., Crovella, M., and Xie, P. (2002). Sampling biases in ip topology measurements. Technical report, Boston University Computer Science, Tech. Rep. BUCS-TR-2002-021.

[Latapy and Pons, 2004] Latapy, M. and Pons, P. (2004). Computing communities in large networks using random walks. *cond-mat/0412568*.

[Latora and Marchiori, 2001] Latora, V. and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87:198701.

[Latora and Marchiori, 2004] Latora, V. and Marchiori, M. (2004). A measure of centrality based on the network efficiency. *cond-mat/0402050*.

[Lawler et al., 1985] Lawler, E. L., Lenstra, J. K., Khan, A. H. G. R., and Shmoys, D. B. (1985). *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley.

[Leland et al., 1995] Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1995). On the selfsimilar nature of ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15.

[Lighthill and Whitham, 1955] Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society A*, 229:317.

[Lopez et al., 2007] Lopez, E., Parshani, R., Cohen, R., Carmi, S., and Havlin, S. (2007). Limited path percolation in complex networks. *cond-mat/070269*.

[Lozano et al., 2007] Lozano, S., Arenas, A., and Sánchez, A. (2007). Mesoscopic structure conditions the emergence of cooperation on social networks. *physics/0612124v2*.

[Lozano et al., 2006] Lozano, S., Duch, J., and Arenas, A. (2006). Community detection in a large social dataset of european projects. In *Sixth SIAM - International Conference on Data Mining*.

[Lusseau et al., 2003] Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, 54:396–40.

[Maerivoet and de Moor, 2005] Maerivoet, S. and de Moor, B. (2005). Cellular automata models of road traffic. *Physics Reports*, 419:1–64.

[Mahadevan et al., 2006] Mahadevan, P., Krioukov, D., Fomenkov, M., Huffaker, B., Dimitropoulos, X., kc claffy, and Vahdat, A. (2006). The internet as-level topology: Three data sources and one definitive metric. *ACM SIGCOMM Computer Communications Review*, 36:17–26.

[Massen and Doye, 2005] Massen, C. P. and Doye, J. P. K. (2005). Identifying communities within energy landscapes. *Phys. Rev. E*, 71:046101.

[Mezard et al., 1987] Mezard, M., Parisi, G., and Virasoro, M. (1987). *Spin Glass Theory and Beyond*. World Scientific Publishing Company.

[Milgram, 1963] Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67:371–378.

[Molloy and Reed, 1995] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Combinatorics, Probability and Computing*, 6:161–179.

[Molloy and Reed, 1998] Molloy, M. and Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7:295–305.

[Moreno et al., 2003] Moreno, Y., Pastor-Satorras, R., Vázquez, A., and Vespignani, A. (2003). Critical load and congestion instabilities in scale-free networks. *Europhys. Lett.*, 62:292.

[Motter, 2004] Motter, A. (2004). Cascade control and defense in complex networks. *Phys. Rev. Lett.*, 93:098701.

[Newman, 2003a] Newman, M. (2003a). A measure of betweenness centrality based on random walks. *Social Networks*, 27:39–54.

[Newman, 2000] Newman, M. E. J. (2000). Models of the small world. *J. Stat. Phys.*, 101:819–841.

[Newman, 2001a] Newman, M. E. J. (2001a). Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(016131).

[Newman, 2001b] Newman, M. E. J. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(016132).

[Newman, 2002] Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701.

[Newman, 2003b] Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

[Newman, 2004a] Newman, M. E. J. (2004a). Analysis of weighted networks. *Phys. Rev. E*, 70:056131.

[Newman, 2004b] Newman, M. E. J. (2004b). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133.

[Newman, 2006a] Newman, M. E. J. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104.

[Newman, 2006b] Newman, M. E. J. (2006b). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, 103:8577–8582.

[Newman and Girvan, 2004] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

[Newman and Leicht, 2007] Newman, M. E. J. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci.*, 104:9564–9569.

[Noh and Rieger, 2004] Noh, J. and Rieger, H. (2004). Random walks on complex networks. *Phys. Rev. Lett.*, 92:118701.

[Norton, 2004] Norton, W. B. (2004). The evolution of the u.s. internet peering ecosystem. Technical report, Equinix White Papers.

[Palla et al., 2005] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818.

[Palla et al., 2007] Palla, G., Farkas, I. J., Pollner, P., Derényi, I., and Vicsek, T. (2007). Directed network modules. *New Journal of Physics*, 9:186.

[Papadimitriou and Steiglitz, 1997] Papadimitriou, C. H. and Steiglitz, K. (1997). *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications.

[Park et al., 1996] Park, K., Kim, G. T., and Crovella, M. E. (1996). On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proceedings of the Fourth International Conference on Network Protocols (ICNP'96)*, pages 171–180.

[Park and Willinger, 2000] Park, K. and Willinger, W. (2000). *Self-Similar Network Traffic and Performance Evaluation*. Wiley-Interscience.

[Pastor-Satorras et al., 2001] Pastor-Satorras, R., Vázquez, A., and Vespignani, A. (2001). Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87:258701.

[Percacci and Vespignani, 2003] Percacci, R. and Vespignani, A. (2003). Scale-free behavior of the internet global performance. *Eur. Phys. J. B*, 32:411–414.

[Pujol et al., 2006] Pujol, J. M., Béjar, J., and Delgado, J. (2006). Clustering algorithm for determining community structure in large networks. *Phys. Rev. E*, 74:016107.

[Radicchi et al., 2004] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci.*, 101(9):2658–2663.

[Ravasz et al., 2002] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

[Reichardt and Bornholdt, 2004] Reichardt, J. and Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a q-state potts model. *Phys. Rev. Lett*, 93:218701.

[Reichardt and Bornholdt, 2006] Reichardt, J. and Bornholdt, S. (2006). When are networks truly modular? *cond-mat/0606220 v1*.

[Rosvall and Bergstrom, 2007] Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci.*, 104:7327–7331.

[Sales-Pardo et al., 2007] Sales-Pardo, M., Guimerà, R., Moreira, A. A., and Amaral, L. A. N. (2007). Extracting the hierarchical organization of complex systems. *Proc. Natl. Acad. Sci*, 104:15224–15229.

[Scott, 2000] Scott, J. (2000). *Social Network Analysis, a handboook*. SAGE publications.

[Seidman, 1983] Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, 5:269–287.

[Serrano et al., 2005] Serrano, M., Boguñá, M., and Díaz-Guilera, A. (2005). Competition and adaptation in an internet evolution model. *Phys. Rev. Lett.*, 94:038701.

[Serrano et al., 2006] Serrano, M., Boguñá, M., and Díaz-Guilera, A. (2006). Modeling the internet. *Eur. Phys. J. B*, 50:249–254.

[Shannon and Weaver, 1963] Shannon, C. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

[Sherrington and Kirkpatrick, 1975] Sherrington, D. and Kirkpatrick, S. (1975). Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35(26):1792–1796.

[Singh and Gupte, 2005] Singh, B. and Gupte, N. (2005). Congestion and decongestion in a communication network. *Phys. Rev. E*, 71:055103(R).

[Sole and Valverde, 2001] Sole, R. and Valverde, S. (2001). Information transfer and phase transitions in a model of internet traffic. *Physica A*, 289:595–695.

[Sreenivasan et al., 2007] Sreenivasan, S., Cohen, R., Lopez, E., Toroczkai, Z., and Stanley, H. E. (2007). Communication bottlenecks in scale-free networks. *Phys. Rev. E*, 75:036105.

[Tadic et al., 2007] Tadic, B., Rodgers, G., and Thurner, S. (2007). Transport on complex networks: Flow, jamming and optimization. *International Journal of Bifurcation and Chaos*, 17.

[Tadic et al., 2004] Tadic, B., Thurner, S., and Rodgers, G. (2004). Traffic on complex networks: Towards understanding global statistical properties from microscopic density fluctuations. *Phys. Rev. E*, 69:036102.

[Taylor, 1961] Taylor, L. (1961). Aggregation, variance and the mean. *Nature*, 189:732–735.

[Valverde and Solé, 2002] Valverde, S. and Solé, R. (2002). Self-organized critical traffic in parallel computer networks. *Physica A*, 312:636–648.

[Vázquez et al., 2002a] Vázquez, A., Pastor-Satorras, R., and Vespignani, A. (2002a). Internet topology at the router and autonomous system level. *cond-mat/0206084*.

[Vázquez et al., 2002b] Vázquez, A., Pastor-Satorras, R., and Vespignani, A. (2002b). Large-scale topological and dynamical properties of the internet. *Phys. Rev. E*, 65:066130.

[Ward, 1063] Ward, J. H. (1063). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 53(301):263–244.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis, methods and applications*. Cambridge University Press.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

[Waxman, 1996] Waxman, B. (1996). Routing of multipoint connections. *IEEE J. Select. Areas Commun*, 6:1617–1622.

[Willinger et al., 2002] Willinger, W., Govindan, R., Jamin, S., Paxson, V., and Shenker, S. (2002). Scaling phenomena in the internet: Critically examining criticality. *Proc. Natl. Acad. Sci*, 99:2573–2580.

[Wu and Huberman, 2004] Wu, F. and Huberman, B. (2004). Finding communities in linear time: a physics approach. *Eur. Phys. J. B*, 38:331–338.

[Yook et al., 2002] Yook, S.-H., Jeong, H., and Barabási, A. L. (2002). Modeling the internet's large-scale topology. *Proc. Natl. Acad. Sci*, 99:13382–13386.

[Zachary, 1977] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33.

[Zegura et al., 1996] Zegura, E., Calvert, K., and Bhattacharjee, S. (1996). How to model an internetwork. *IEEE Infocom*, 2:594–602.

[Zhao et al., 2005] Zhao, L., La, Y., Park, K., and Ye, N. (2005). Onset of traffic congestion in complex networks. *Phys. Rev. E*, 71:026125.

[Zhou, 2003a] Zhou, H. (2003a). Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67:061901.

[Zhou, 2003b] Zhou, H. (2003b). Network landsape from a brownian particle's perspective. *Phys. Rev. E*, 67:041908.

[Zhou and Lipowsky, 2004] Zhou, H. and Lipowsky, R. (2004). Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. *Lecture Notes in Computer Science*.

[Zhou and Lipowsky, 2005] Zhou, H. and Lipowsky, R. (2005). The yeast protein-protein interaction map is a highly modular network with a staircase community structure. *preprint*.

[Zhou and Mondragon, 2003] Zhou, S. and Mondragon, R. J. (2003). Towards modelling the internet topology - the interactive growth model. In *Proc. of the 18th International Teletraffic Congress*.