Probabilistic Models of Growth of Networks

Batagelj, Vladimir
University of Ljubljana, FMF, Department of Mathematics
Jadranska 19
1000 Ljubljana, Slovenia
E-mail: vladimir.batagelj@fmf.uni-lj.si

Kejžar, Nataša
University of Ljubljana, MF, IBMI
Vrazov trg 2
1000 Ljubljana, Slovenia
E-mail: natasa.kejzar@mf.uni-lj.si

The development of Internet and recent studies of complex systems have significantly increased the interest in modeling classes of graphs and networks. While random graphs have been studied for a long time, the standard (Erdős and Rényi; Gilbert) models appear to be inappropriate because they do not share the characteristics observed in real life systems. Several new models were proposed that better match the reality. Most of them are variations of the the preferential attachment model (Barabási and Albert). Often we are also interested in random graphs/networks with some additional properties (connectivity, planarity, ...) or resulting from the evolution under certain rules. The probabilistic inductive classes of graphs allow us to consider also these additional requirements.

Classical random graphs and scale-free networks

The notion of random graph was introduced in 1959 by Paul Erdős and Alfréd Rényi [14] and Edgar Gilbert [16]. In Gilbert's approach the set $G(n,p)$ consists of random graphs in which each pair of n nodes is linked with a given probability p. In Erdős–Rényi's approach the set $G(n,m)$ consists of all graphs with n nodes and m edges with uniform probability $\binom{m}{n}^{-1}$. It turns out that in most applications both notions of random graphs are practically interchangable, provided that $n \approx pm$. We shall call these models of random graphs classical.

The theory of classical random graphs is well developed (see Bollobás [6]). Their degree distribution is binomial (in the limit Poisson's) and most of the nodes have degree (very) close to the average degree. The graph property Q is *monotone* iff for it holds: let graph H be a subgraph of graph G and H has the propery Q then also graph G has the property Q. An important observation of Erdős and Rényi was that for many monotone graph properties a threshold value α of p exists such that for graphs with $p < \alpha$ it is very unlikely that a graph has that property, and most of graphs with $p > \alpha$ have that property – a kind of phase transition. For example [29]:

- for $p > n^{-3/2}$ edges with common node appear;
- for $p > \frac{1}{n}$ triangles and other cycles appear in the graph, the graph becomes nonplanar, and soon appears also the *giant component*;
- for $p > \frac{\ln n}{n}$ almost all graphs are connected.

Real-life networks are usually not random in the classical sense. The analysis of their degree distributions in late 90-ties by researchers from University of Notre Dame showed that most of them are very far from the normal [2]. Usually in real life networks there are many nodes of low degree and also some nodes with very high degree (heavy tail) – the distribution is not concentrated around the mean value.

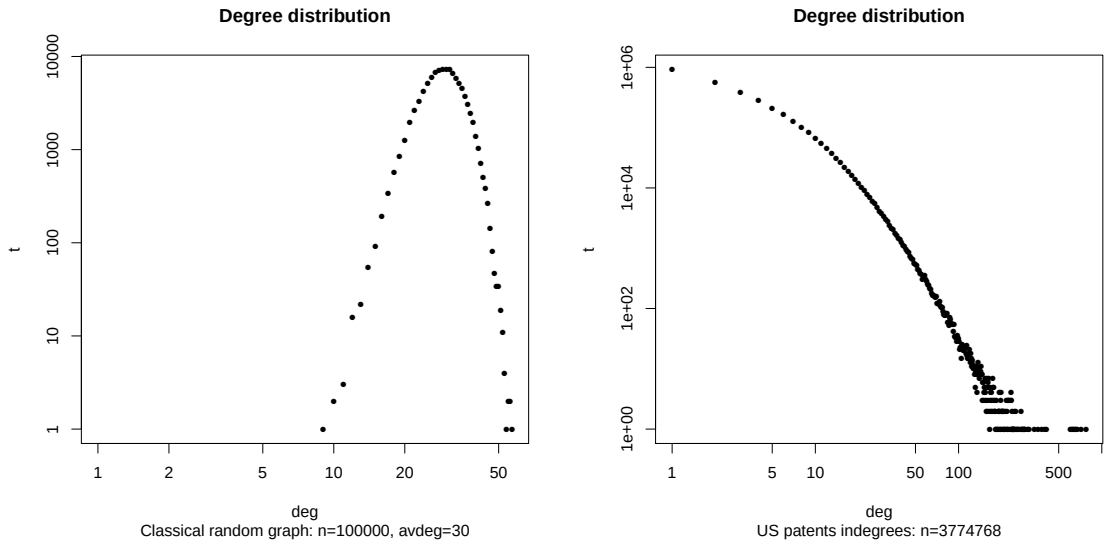

Figure 1: Degree distributions: classical random graph and US patents network

As an example, on the left side of Figure 1 a degree distribution of classical random graph on $n = 100000$ nodes with average degree $\overline{\deg} = 30$ is presented. On the right side a degree distribution for US Patents citation network ($n = 3774768$ and $m = 16522438$) is presented. Both distributions are displayed in log-log scale. Evidently they are very different.

Somehow surprisingly many degree distributions of real-life networks are presented in log-log scale with a curve close to a line [2] – the distribution is close to the *power law* – the probability p_d that a node has a degree d equals to $p_d = cd^{-\gamma}$. Since for this distribution no meaningful selection of the scale exists this type of networks was named *scale-free*. For a detailed discussion about the notion of scale-free network see Li et al. [23]. For methods for computing the parameter γ see Newman [26].

Albert-László Barabási [2] also provided the first explanation of the reasons for the power law distribution. He presented a model of growing network to which new nodes are added and linked to the already existing nodes following the *preferential attachment* rule: the node is linked to an old node with the probability proportional to the degree of the old node. For these networks we can show that the average length of geodesics (shortest paths) is $O(\log n)$, and that they are resilient against random node or edge removals (random attacks), but soon become disconnected when large degree nodes (Achilles' heel) are removed (targeted attacks).

Several alternative models and improvements to produce scale-free (like) networks with some additional properties encountered in real-life networks were proposed in the following years: copying (Kleinberg [20]), combining random and preferential attachment (Pennock et al. [28]), R-mat (Chakrabarti et al. [8]), aging, fitness, nonlinear preferences, and others.

The scale-free networks theory was applied for solving problems in different fields such as: searching in networks (Adamic et al. [1]) and spreading of epidemies or innovations (Barthélemy, Barrat, Pastor-Sattoras, Vespignani [9]). For detailed overviews of results on scale-free networks see Dorogovtsev and Mendes [11], and Newman at al. [25], Kolaczyk [21], and Easley and Kleinberg [12].

Generating large sparse random graphs and networks

For studying and predicting behaviour of Internet, communication, transportation, biological, social and other types of networks the simulation approach is usually used. As a support for it we need efficient generators of networks of the corresponding types [7, 15].

Large networks are usually (very) sparse. In most real-life networks the capacity of a node to maintain links with others is limited. In sociology such a bound is known as the Dunbar's number (Hill and Dunbar

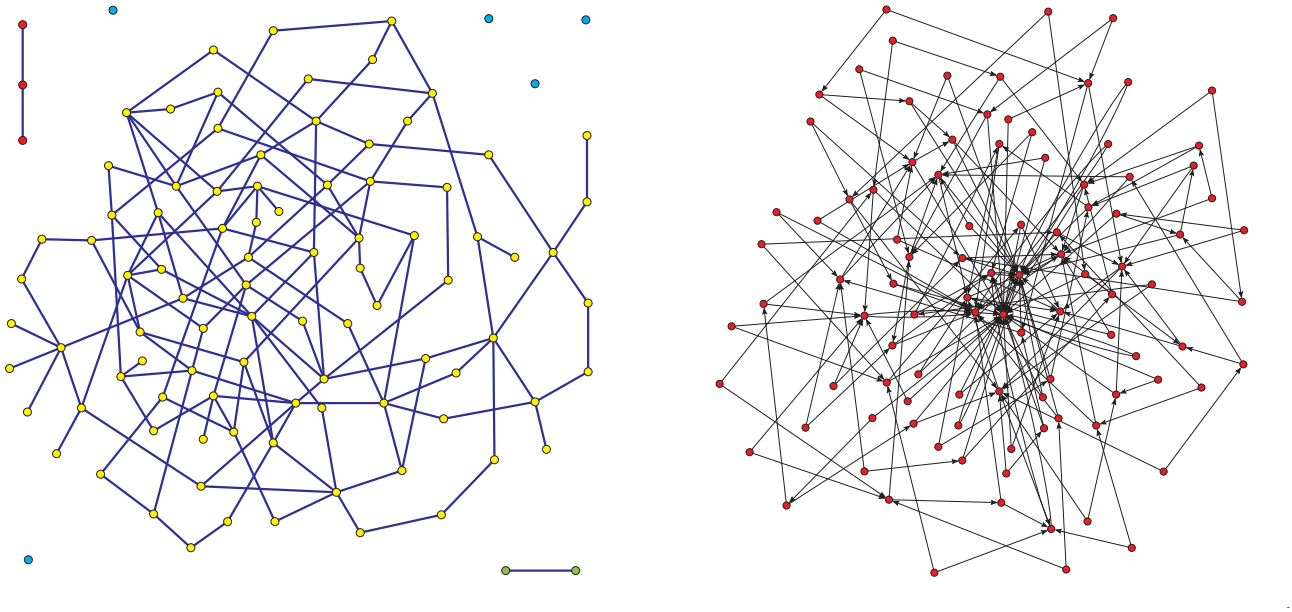

Figure 2: Random graph of Gilbert's type, $n = 100$, $\overline{\deg} = 3$; and scale-free random graph, $n = 100$, $d = 2$

[18]). Usually it is approximated by 150. Therefore the average degree in the network is not large. For this reason the standard (based directly on the definitions) algorithms for generating random graphs of selected type can be inefficient. In Batagelj and Brandes [5] we presented several fast algorithms for generating large sparse random graphs and networks of different types.

For example, the generation of random graphs of Gilbert's type is equivalent to the filling of lower triangle of graph's adjacency matrix with Bernoulli sequence with parameter p of length $\binom{n}{2}$. In generating large and sparse such graphs we can replace it with puting the value 1 in positions determined by the corresponding geometrically distributed steps. This gives us the following much faster generator.

```
Gilbert <-
# generates a random undirected graph of Gilbert's type
# on n nodes with expected average degree ad and writes
# it on the file fnet in Pajek's format.
# based on ALG.1 from: V. Batagelj, U. Brandes:
#   Efficient generation of large random networks
function(fnet,n,ad){
  net <- file(fnet,"w"); cat("*nodes",n,
    "\n% random Gilbert's graph / n = ",n," ad = ",ad,"\n*edges\n",file=net)
  logQ <- log(1-ad/(n-1));  v <- 1;  w <- -1
  while (v < n){
    w <- w + 1 + trunc(log(1-runif(1,0,1))/logQ)
    while (w >= v) {w <- w-v; v <- v+1}
    if (v < n) cat(v+1,w+1,'\n',file=net)
  }
  close(net)
}

Gilbert("gilbert.net",100,3.0)
```

Since in large sparse networks the probability p is very small in the R function `Gilbert` the parameter `p` is replaced by a more intuitive average degree `ad` and computed internally using the relation

$$\overline{\deg} = p \cdot (n-1)$$

Similary, representing edges with pairs of nodes and observing that the number of copies of a node in a table equals its degree, we get the following fast generator of scale-free graphs.

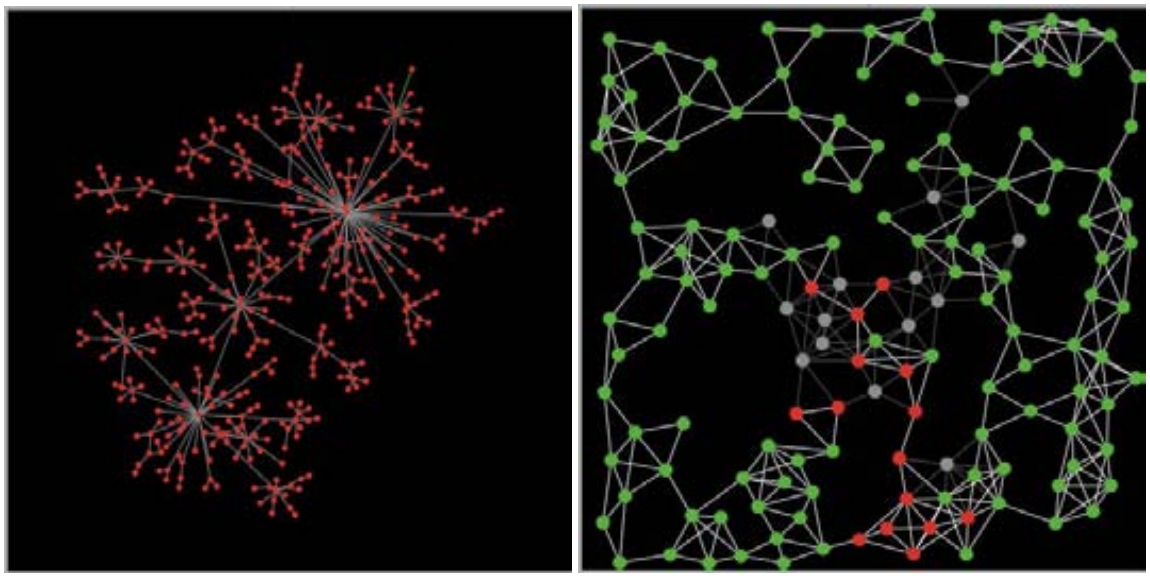

Figure 3: Netlogo: evolution of scale-free network and spreading of the virus

```
dice <- function(n=6){return(1+trunc(n*runif(1,0,1)))}

ScaleFreeBrstNet <-
# generates a random directed scale free graph
# on n nodes with d attachments to existing nodes
# and stores it on the file  fnet  in Pajek's format
# based on ALG.5 from: V. Batagelj, U. Brandes:
#   Efficient generation of large random networks
function(fnet,n,d){
  net <- file(fnet,"w"); cat("*nodes",n,"\n",file=net)
  k <- 0; m <- n*d; L <- rep(0,2*m)
  cat('% random scale free graph / n = ',n,' d = ',d,'\n',file=net)
  for(v in 1:n) for (i in 1:d) {
    k <- k+1; L[[k]] <- v; r <- dice(k);
    k <- k+1; L[[k]] <- L[[r]]
  }
  cat("*arcs\n",file=net)
  for (i in 1:m) cat(L[[2*i-1]],L[[2*i]],'\n',file=net)
  close(net)
}

ScaleFreeBrstNet("scaleFree.net",100,2)
```

Examples of graphs generated with both algorithms are displayed in Figure 2.

Probabilistic inductive classes of graphs

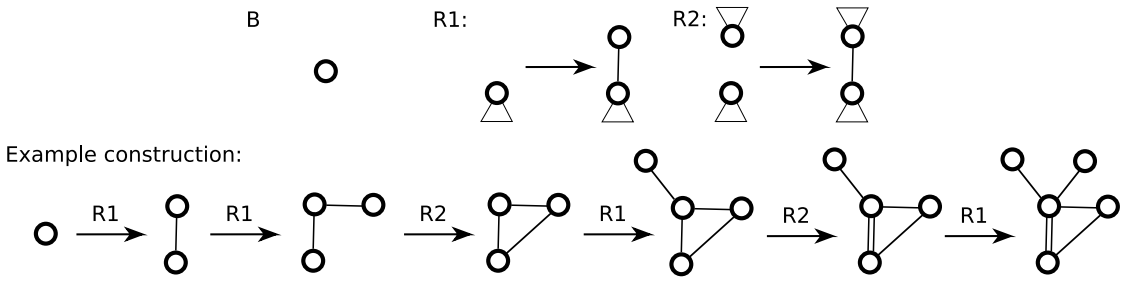

Figure 4: Basic graphs and rules

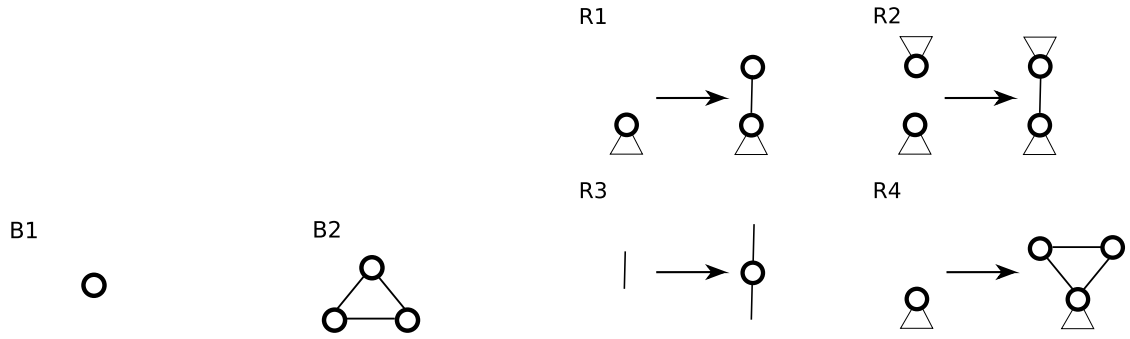

Figure 5: Basic graphs and rules

A random network can be viewed also as a result of an evolution process starting from some simple network in which next network is obtained from the current network using some (local) transformation. Some such models are implemented in programming language Netlogo [31] (see Figure 3). The class of graphs/networks that can be obtained in this way can be described using inductive definitions (Curry [10], Batagelj [4]) or in more formalized setting using graph grammars (Ehrig et al. 1991 [13]). In our research we prefer the less formal inductive definitions because they are easier to adapt to specific characteristics of the classes of our interest.

The notion of inductive class of graphs can be extended by assigning probabilities to events in the evolution process. In the paper (Kejžar et al. [19]) we presented the following definition:

A *probabilistic inductive class of graphs* (*PICG*), $\mathcal{I}$, is given by:

1. class $\mathcal{B}$ of initial graphs, the *basis* of PICG,
2. probability distribution specifying how the initial graph is chosen from class $\mathcal{B}$,
3. class $\mathcal{R}$ of generating rules, each with distinguished *left element* to which the rule is applied to replace it with the *right element*,
4. probability distribution specifying how the rules from class $\mathcal{R}$ are applied, and, finally,
5. a set of probability distributions specifying how the left elements for every rule in class $\mathcal{R}$ are chosen.

A random graph is obtained by starting from some randomly selected basic graph from the basis $\mathcal{B}$ and applying on it a randomly selected generating rules from $\mathcal{R}$ on randomly selected subgraph isomorphic to the rule's left element. On the so obtained graph the next randomly selected rule is applied, and so on. The PICG $\mathcal{I}$ consists exactly of graphs that can be obtained in this way in a finite number of steps. The sequence of graphs corresponding to these steps, enriched with the information about the applied rule, is called the *construction sequence* of a graph from the class.

In Figure 5 a simple ICG $\mathcal{I}(B; R1, R2)$ and an example of construction sequence are presented.

For the base graphs and rules from Figure 5 the ICG $\mathcal{I}(B1; R1, R2)$ is the class of all connected (undirected) graphs, $\mathcal{I}(B2; R2, R3)$ is the class of all 2-node-connected graphs, and $\mathcal{I}(B2; R2, R3, R4)$ is the class of all 2-edge-connected graphs.

In our paper [19] we analyzed these three inductive definitions for the case when the generating rules have constant probabilities to be selected and the left part subgraph is selected with the uniform probability among available isomorphic subgraphs. For such relatively simple definitions theoretical answers to some questions can be obtained using the mean-field approach from theoretical physics. For more complicated definitions it seems that the only way to get some approximate answers is the simulation approach.

The number of classes that can be described as PICGs depends on the limitations we impose on the language for expressing the rules. In general we can allow also parametrized schemes of rules that produce

rules only after specifying the values of parameters – they are finitely describing possibly infinite sets of rules. If needed we can also introduce a precedence among groups of rules – the rules with lower precedence are applied only when no rule with higher precedence can be applied.

In networks the graph structure is enriched by values in nodes and/or on links. Often they can be treated as colors. Taking a given network as a base network and introducing the rules that change colors, we can use PICGs also for studying different processes on networks – for example balancing in the signed networks [17].

An interesting question to be solved is also how to estimate the probabilities of generating rules from the realized graphs/networks.

REFERENCES

[1] Adamic, L.A., Lukose, R.M., Huberman, B.A.(2002) *Local Search in Unstructured Networks*. in Handbook of Graphs and Networks: From the Genome to the Internet, S. Bornholdt, and H.G. Schuster (eds.), Wiley-VCH, Berlin.

[2] Barabási, A-L. and Albert, R. (1999) "Emergence of scaling in random networks", Science, 286:509-512.

[3] Albert, R. and Barabási, A-L. (2002) *Statistical mechanics of complex networks*. Reviews of Modern Physics, Vol. 74, 47-97.

[4] Batagelj, V. (1985). Inductive classes of graphs. Proceedings of the Sixth Yugoslav Seminar on Graph Theory, Dubrovnik, 4356. URL: `http://vlado.fmf.unilj.si/vlado/projects/indcla.htm`

[5] Batagelj, V. and Brandes, U. (2005) *Efficient Generation of Large Random Networks*. Physical Review E 71, 036113, 2005.

[6] Bollobás, B. (2001) *Random Graphs*. Cambridge University Press.

[7] BRITE: Boston University Representative Internet Topology Generator, URL: `http://cs-www.bu.edu/brite/`

[8] Chakrabarti, D., Zhan, Y., Faloutsos, C. (2004) *R-MAT: A Recursive Model for Graph Mining*. in SIAM Data Mining 2004, Orlando, Florida, USA.

[9] Complex Networks Collaboratory: `http://cxnets.googlepages.com/`

[10] Curry, H.B. (1963) Foundations of mathematical logic. McGraw-Hill, New York.

[11] Dorogovtsev, S.N., Mendes, J.F.F. (2003) *Evolution of Networks: From Biological Nets to the Internet and Www*. Oxford University Press.

[12] Easley, D. and Kleinberg, J. (2010) Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.

[13] Ehrig, H., Kreowski, H. J., Rozenberg, G., (eds.) (1991). Graph Grammars and Their Application to Computer Science, Lecture Notes in Computer Science, vol. 532. Berlin and Heidelberg: Springer-Verlag.

[14] Erdős, P. and Rényi, A. (1959). "On Random Graphs. I.". Publicationes Mathematicae 6: 290297.

[15] GT-ITM: Georgia Tech Internetwork Topology Models,
URL: `http://www.cc.gatech.edu/fac/Ellen.Zegura/graphs.html`

[16] Gilbert, E.N. (1959) Random Graphs. Annals of Mathematical Statistics. Volume 30, Number 4 , 1141-1144.

[17] Heider, F. (1946). Attitudes and cognitive organization. Journal of Psychology, 21, 107112.

[18] Hill, R.A. and Dunbar, R.I.M. (2002) Social network size in humans, Human Nature, 14(1): 5372.

[19] Kejžar, N., Nikoloski, Z., Batagelj, V.: Probabilistic Inductive Classes of Graphs. Journal of Mathematical Sociology 32: 85-109, 2008.

[20] Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999) *The Web as a graph: measurements, models and methods*. Proceedings of the 5th International Computing and combinatorics Conference.

[21] Kolaczyk, E.D. *Statistical Analysis of Network Data: Methods and Models*. Springer, Berlin 2009.

[22] Leskovec, J., Kleinberg, J., Faloutsos, C. (2006) *Laws of Graph Evolution: Densification and Shrinking Diameters*.

[23] Li, L., Alderson, D., Tanaka, R., Doyle, J.C., Willinger, W. (2005) *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications*. cond-mat/0501169.

[24] Moreno, Y., Nekovee, M., and Vespignani, A. (2004). Efficiency and reliability of epidemic data dissemination in complex networks. Physical Review E, 69, 055101.

[25] Newman, M.E.J., Barabási, A-L., Watts, D. (2006) *The Structure and Dynamics of Networks*. Princeton Studies in Complexity.

[26] Newman, M.E.J. (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46:323–351.

[27] de Nooy, W., Mrvar, A., Batagelj, V. (2005) *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

[28] Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., Giles, C.L. (2002) *Winners dont take all: Characterizing the competition for links on the web*. PNAS 99(8), 52075211.

[29] Spencer, J. (2001) The Strange Logic of Random Graphs. Springer.

[30] Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks Nature 393, 440-442.

[31] Wilensky, U. (1999) Netlogo. Center for Connected Learning and Computer-Based Modeling. Northwestern University, Evanston, IL.