Kernel Models for Complex Networks

Milena Mihail Yorgos Amanatidis Stephen Young

School of Computer Science and School of Mathematics
Georgia Institute of Technology

Abstract: We advocate the study of graph models for complex networks based on kernel functions over metric spaces. These are hybrids between: (a)macroscopic random graph models for complex networks (b)microscopic models that incorporate network semantics, reminiscent of machine learning and information retrieval primitives. In particular, kernel-based models assign explicit semantics to network nodes and links. These semantics can capture fundumental properties, such as hierarchy and clustering; the latter arise repeatedly in real complex networks, but they do not arise in pure random graphs. At the same time, kernel-based models maintain the conceptual, analytical and implementational clarity of random graph models.

1 Introduction

Complex networks arise naturally in societies, economies and technologies. In recent years, the study of complex networks has intensified, due to the dramatic growth of the WWW, the continuous emergence of novel network applications running over the WWW, and the capability of current technology to collect and store data arising from complex networks. Models for complex networks are fundamental for network understanding, prediction and simulation. Such models should be consistent with general network characteristics. In addition, such models should capture further detailed network characteristics that are important to the particular application. These further characteristics may vary substantially accross different applications. If the application calls for repeated generation of synthetic network instances (as would typically be the case, for example, in scaling studies), it is also important that instances can be generated very efficiently.

Two nearly universally obseved characteristics of complex networks, namely heavy tailed degree distributions and the small world phenomenon, have given rise to network models vastly different from the 50 year old classical Erdös-Renyi random graphs [7]. At the same time, Kleinberg's pioneering work on navigability [10] pointed out that additional critical parameters should be involved in network studies; the range of such parameters can determine network function.

The first generation of complex network models, developped by physicists, mathematicians, computer scientists and economists [5, 2, 6, 14, 7, 9], can be broadly categorized as macroscopic and microscopic. Macroscopic models are random graphs with skewed degree distributions and possessing the small world property (we skip further studies characteristics from this short paper). The advantage of macroscopic models is conceptual simplicity, amenability to mathematical analysis, and relatively efficient algorithms generating synthetic network instances. Their disadvantage is that they are missing network semantics. Consequently, such models have been noted to fail, especially in cases where network elements have explicit additional semantics.

Two fundamental cases where macroscopic random graph models have been noted to fail involve hierarchy and clustering. These notions have immediate intuitive understanding and invoke semantics explicitly. Hierarchy speaks of distinct classes of significance. Clustering speaks of distinct preferences in forming associations. Local clustering involves local preferences, such as "the friend of my friend is more likely to be my friend". Global clustering involves very well connected communities, indicating collective and very specific common interests. It has been noted extensively that pure random graph models fail to captute both hierarchy and clustering [13, 12, 11, 8]. And yet, for every application involving a complex

network where hierarchy and/or clustering is present, it is particularly important that a corresponding network model should capture this property.

Kernel random graphs is a new genre of complex models, which endows usual random graphs with semantics on their nodes and their links. The significance of kernel random graphs is that nodes and links have semantics. Node semantics are expressed by representing nodes as vectors in high dimenstional spaces. These nodes are distributed according do a distribution μ. The links of the graph are determined by the the semantics of nodes at the endpoint of each link. In particular way by which node semantics determine how (with what probability) each pair of nodes is connected with a link, is expressed by a so-called kernel function κ, hence the name of the model. Kernel random graphs become powerful when properties of the generated graphs can be well characterized in terms of the new "parameters" μ and κ. The aim is to capture classes of μ's and κ's which provably result in properties that are observed in real complex networks. In Section 2 we describe kernel random graphs more formally, and report recent results suggesting that they are amenable to analysis. In Sections 3 and 4 we outline the potential of kernel random graphs to capture hierarchy and clustering.

2 Random Graph Models based on Kernel Functions

The Model $\mathcal{G}_\mu^{\kappa,g}(n)$

Kernel random graphs is a new genre of complex network models, which endows usual random graphs with semantics on their nodes and their links [3, 4]. For any given number of network nodes n, a kernel random graph model generates a graph on n nodes.

Semantics on nodes are captured by representing nodes as vectors in d-dimensional space, one dimension for each relevant node attribute. For example, think of each attribute as a distinct characteristic, or interest. Each network node is determined by its values on each attribute (extend of characteristic, amount of interest). d is fixed and independent of n. From the application point of view, d being fixed is a reasonable starting point. Each node can be though of as a local entity with limited resources, independent of the size of the entire population. Kernel random graphs have been defined for more general metric spaces in [3], but in this short paper we restrict the exposition to d-dimensional space.

Nodes are sampled from a fixed, bounded but otherwise general distribution $\mu \in \Re^d$. The generality of μ can capture a wide range of statistical behaviors. The condition that μ is bounded is entirely reasonable. It says that, for each network node, the total sum of the values of the node on each attribute is bounded. Clearly, a node with bounded resources cannot keep track of of ever growing levels of characteristics or interests. The condition that μ is fixed and independent of n is very strong. It says that, statistically, the overall distribution of emantical characteristics of nodes does not change. On the other hand, we know that, even in networks with bounded local resources, there exist local network dynamics that can change the distribution of network node characteristics over time or as the network grows. In the end of Section 4 we shall see a case where the condition on μ is relaxed.

Semantics on links are captured by a kernel function κ, which maps pairs of nodes to the probability that they form a link. The conditions on κ are very mild and entirely natural (should be continious almost everywhere). Finally, we have a function $g = g(n)$ to control the density (average degree, or total number of links) of the generated graph.

A kernel random graph can be then generated as follows: the nodes are generated according to μ and they are d-dim vectors $\vec{x_i}, 1 \leq i \leq n$. A link between each pair of nodes $\vec{x_i}$ and $\vec{x_j}$ is added with probability proportional to $\kappa(\vec{x_i}, \vec{x_j})/g(n)$. We call the class of random graphs, determined by μ, κ, g and n, $\mathcal{G}_\mu^{\kappa,g}(n)$,

In addition to introducing semantics, realize another advantage of kernel random graph models: they explicitly separate the issue of inferring μ and κ, from the issue of characterizing the structural and

functional properties of $\mathcal{G}_{\mu}^{\kappa,g}(n)$. The former is the object of statistical inferencing and learning theory, while the latter is the object of graph theory. Clearly, web science should involve statistical inferencing, learning theory and graph theory.

In particular, for graph theory and web science, the question then becomes: What functions μ and κ give rise to well characterized classes of graphs $\mathcal{G}_{\mu}^{\kappa,g}(n)$? We stress that μ and κ are the new random graph "parameters". They are the precise parameters capturing network semantics. Therefore, properties of $\mathcal{G}_{\mu}^{\kappa,g}(n)$ should be explicitly invoking μ and κ.

Analytical Results for $\mathcal{G}_{\mu}^{\kappa,g}(n)$

Towards chacterizing properties of $\mathcal{G}_{\mu}^{\kappa,g}(n)$ for general classes and in terms of μ and κ, there are two independent lines of work with positive results: (1)Bollobás et al [3] showed that essentially all known sparse inhomogenious random graph models (heavy tailed degree distributions but constant average degree) can be expressed in terms of suitably chosen μ and κ (and more general technical metric spaces). In addition, for sparse $\mathcal{G}_{\mu}^{\kappa,g}(n)$, [3] obtain structural characteristics in terms of μ and κ, for fairly general μ and κ, and under mild assumptions. The main point of [3] is technical and very strong: they demonstrate that mathematical methods previously used for Erdös-Renyi random graphs carry over to much more general classes of graphs. (2)Young and collaborators [17] studied the case where κ is the inner product function. Realize that the inner product is a very natural way to capture similarity. It is used to express similarity throughout machine learning [16, 15]. For $\mathcal{G}_{\mu}^{<\cdot,\cdot>,g}(n)$ with average degree ranging from $\Omega(\log n)$ to $O(n)$, and under no assumptions, [17] express diameter, degree distribution and clustering explicitly and in closed form in terms of μ. They also show that suitable μ's can capture all three hallmarks of complex networks: skewed degrees, low diameter, and local clustering.

Another important advantage of $\mathcal{G}_{\mu}^{\kappa,g}(n)$ is that it easy to simulate. In most cases, each vector x_i can be generated from μ in time $O(\log n)$, and the links $\vec{x_i} \sim \vec{x_j}$ require, in the worst case, $O(n^2)$ experiments; the latter can be greatly improved in many specific cases (e.g. inner product graphs can be generated intimeproportional to their size, which is optimal). Finally, for many classes of κ, including the inner product, a graph over n nodes $\mathcal{G}_{\mu}^{\kappa,g}(n)$ can be made to evolve naturally and very efficiently to a graph over N nodes $\mathcal{G}_{\mu}^{<\cdot,\cdot>,g}(N)$, for $N >> n$.

3 The Case of Hierarchy

Hierarchy is rather well understood intiutively, and is therefore invoked often in network studies. For networks with heavy tailed statistics, hierarchy has been captured by a quantity called "assortativity" [13, 12]. Assortativity involves the sum $\sum_{u\sim v} d_u d_v$, that is, the sum, over all network links $u \sim v$, of the product of the degrees of the nodes u and v. In random graphs, high degree nodes tend to connect to other high degree nodes, and appear to be at the center of the network; in such networks the assortativity $\sum_{u\sim v} d_u d_v$ is large. On the other hand, there are technological networks, such as the intranet of big ISP providers, where nodes are routers with vastly different bandwidths. As expected, high bandwith routers are placed at the center of the network, and are interconnected in relatively sparse regular patterns. At the same time, very high degree nodes correspond to much lower bandwidth routers which split the signal manyways towards the end users at the network periphery. In networking, this is characterized as hierarchy, and the assortativity $\sum_{u\sim v} d_u d_v$ is small. Microscopic models have tried to capture the case of networks with small assortativity by expressing network formation as the result of cost-benefit optimization [12]. As expected, these models are successful for the particular applications. However, the involved combinatorial optimization problems are notoriously hard, and do not scale. Quite interestingly, small assortativity has been also observed in the human gene-protein interaction network, while, for example, the yeast gene-protein interaction

network has much larger assortativity [13, 8]. It appears very unlikely that any process involving cost-benefit combinatorial optimization will capture or distinguish gene-protein interaction networks. Indeed, the only known processes driving the evolution of such networks are population statistics and probabilistic primitives invoked in natural selection. We therefore prefer to have general macroscopic random models, with additional parameters.

Recently, Amanatidis and Mihail [1] developped hierarchical networks, including the case of small assortativity, low degree routers in the core and high degree routers in the periphery, as manifested in networking [12], in $\mathcal{G}_{\mu}^{\kappa,g}(n)$. They used 3-dimensional space. Two dimensions reprent geographic coordinates, as usual. The third dimension explicitly respresents router bandwidth (small, medium or large). Heavy tailed statistics follow by suitable choice of μ: Small bandwidth routers are distributed according to heavy tailed probability distributions in geographic coordinates, as would be the case of end-users. Medium and large routers are distributed uniformly. The kernel function dictates that routers are connected according to hierarchy and geography, with hierarchy assuming precedence.

4 The Case of Clustering

The second case of discrepancy of random graph models from data arising in real complex networks concerns the case of local or global clustering.

Local clustering is the property that $\Pr(u \sim v | u \sim x, v \sim x) > \Pr(u \sim v)$. This property is observed in nearly all real complex networks, and nearly no pure random graph models. For $\mathcal{G}_{\mu}^{<\cdot,\cdot>,g}(n)$, Young [17] derived a closed formula for $\Pr(u \sim v | u \sim x, v \sim x) - \Pr(u \sim v)$, explicitly in terms of properties μ. He showed positive clustering in all cases, except when μ is concentrated in a single point (pure Erdös-Renyi graphs).

Global clustering concerns the existence of much deeper or sparser cuts in real complex networks (much better intraconnected communities), as compared to random graphs [11]. Moreover, the few sparse cuts that occur in random graphs involve small number of nodes. Real complex networks appear to have sparse cuts with substantially larger number of nodes (stronger communities, of bigger size). Mathematically, cut sparsity is measured by a quantity ϕ called "conductance". Deep, or sparse cuts correspond to subsets involving $k = k(n)$ nodes, and conductance $\Phi(k, n)$. In pure random graphs on n nodes, all cuts have conductunce much larger than, roughly, $1/\log n$. In fact, the only cuts that have conductance close to $1/\log n$ involve no more than, roughly $\log n$ nodes. If $\Phi(k, n)$ is known, then it is very interesting to explore how sparse cuts with desired conductance and explicit semantics can be generated in $\mathcal{G}_{\mu}^{\kappa,g}(n)$, if we allow μ and κ to depend on n [4]. It also very interesting to explore what is the precise nature of $\Phi(k, n)$ in [11], and what are the semantics μ of the corresponding cuts. We shall report results in this direction in the full paper.

Acknowledgements

This work was supported by NSF-CCF-TF-0830683, by an NSF-VIGRE Fellowship, and by ACO program and ARC center at Georgia Tech graduate student fellowships.

References

[1] Amanatidis, Y., Green, B., and Mihail, M., Flexible Models for Complex Networks, Center for Algorithms, Randomness and Computation, October 21, 2008.

[2] Barabasi, A., Linked: The New Science of Networks, NY Perseus, 2002.

[3] Bollobás, B., Janson, S. and Riordan, O., The Phase Transition in Inhomogenious Random Graphs, Random Structures and Algorithms, (31), 2007.

[4] Bollobás, B., Janson, S. and Riordan, O., Sparse Random Graphs with Clustering, preprint, August 2008.

[5] Bornholdt, S. and Schuster, H., Handbook on Graphs and Networks: from the Genome to the Internet, Wiley, 2002.

[6] Dorogovetisev, S. N. and Mendes, J. F. F., Evolution of Networks: from Biological Nets to the Internet and WWW, Oxford University Press, 2003.

[7] Durrett, R., Random Graph Dynamics, Cambridge University Press, 2006.

[8] Hallinan, J. Gene Duplication and Hierarchical Modularity in Intracellular Interaction Networks, BioSystems (74), 2004.

[9] Jackson, M., Economic and Social Networks, Princeton University Press, 2008.

[10] Kleinberg, J., Complex Networks and Decentralized Search Algorithms, Proceedings of the International Congress of Mathematicians, 2006.

[11] Leskovec, J., Lang, K., Dasgupta, A. and Mahoney, M., Community Structure in Large Networks: Natural Cluster Sizes and the Absense of Large Well Defined Clusters, preliminary version in WWW08.

[12] Li, L., Alderson, D., Willinger, W., and Doyle, J., A First-Principles Approach to Understanding the Internet's Router Level Topology, Proceedings ACM-Sigcomm, 2004.

[13] Newman, M., Assortative Mixing in Networks, Phys. Rev. Lett. 89, 2002.

[14] Newman, M. and Barabasi A.L. and Watts, D., The Structure and Dynamics of Networks, Princeton University Press, 2006.

[15] Shawe-Taylor, J. and Christianini, N., Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[16] Bernhard Schölkopf, B. and Smola, A.J., Learning with Kernels, MIT Press, 2002.

[17] Young, S., Random Dot Product Graphs as Flexible Models for Complex Networks, Ph.D. Thesis, Georgia Institute of Technology, 2008. http://www.math.gatech.edu/young/research/papers/thesis.pdf