

# Models of Real-World Random Networks

## Abstracts for Lectures

SPEAKER: Aaron Clauset

TITLE: On the Bias of Traceroute Sampling, or: Why almost every network looks like it has a power law

Understanding the structure of the Internet graph is a crucial step for building accurate network models and designing efficient algorithms for Internet applications. Yet, obtaining its graph structure is a surprisingly difficult task, as edges cannot be explicitly queried. Instead, empirical studies rely on traceroutes to build what are basically single-source, all-destinations, shortest-path trees. These trees only sample a fraction of the network's edges, and a recent paper by Lakhina et al. found empirically that the sample is intrinsically biased. For instance, the observed degree distribution under traceroute sampling exhibits a power law even when the underlying degree distribution is Poisson.

In this talk, we will explore the traceroute sampling bias systematically, and, for a very general class of underlying degree distributions, calculate the observed distributions explicitly. To do this, we use a careful, continuous-time realization of the process of exposing the BFS tree of a random graph, and show that the observed distributions are in fact sharply concentrated. As example applications of our machinery, we show how traceroutes find power law degree distributions in both  $d$ -regular and Poisson-distributed random graphs. Thus, our work puts the observations of Lakhina et al. on a rigorous footing, and extends it to nearly arbitrary degree distributions.

-----

SPEAKER: Anthony Bonato

TITLE: Infinite limits and models of the web graph

Abstract: The existing research on models for the web graph deals almost exclusively with large finite graphs. However, in the natural sciences, models are often studied by taking the infinite limit. Limiting behaviour can be a powerful tool in clarifying the similarities and differences between models. The first study of limiting behaviour of web graph models was made by Jeannette Janssen and the SPEAKER, who made a study of deterministic infinite graphs satisfying the locally e.c. adjacency property (originally called property (B)). As motivation, this adjacency property is satisfied with probability 1 by limit graphs generated by the copying model. Subsequent work on limiting behaviour of preferential attachment web graph models was done by Jon and Robert Kleinberg.

We present some new research on infinite limits and web graph models, focussing on a new generalized copying model introduced by Bonato and Janssen. This model incorporates design elements from copying models of the web graph, and the duplication model for biological networks introduced by Fan Chung and her co-authors. We will discuss adjacency properties satisfied with probability 1 by the limits of graphs generated by the

model, and explain how these properties can distinguish choices for the parameters of the model. The limits of graphs generated by the model behave in some ways like the infinite random graph. For example, they display rich symmetry as realized by an abundance of automorphisms. To this end, we will describe work done with Peter Cameron on the automorphism groups and endomorphism monoids of the limits.

-----

SPEAKER: Ashish Goel

TITLE: Sharp thresholds in geometric random graphs, with algorithmic implications

Geometric random graphs are a widely used model of sensor networks. We will prove that all monotone properties have sharp thresholds in these networks. We will also show that these networks have a nice containment property as the communication radius increases. We will point out implications for the mixing time of random walks and hence the convergence time of gossip algorithms on such networks.

-----

SPEAKER: Kevin McCurley

TITLE: Hierarchical structure in real world networks.

Random graph models tend to be designed with two purposes in mind: enough simplicity to allow mathematical analysis, and enough detail to describe some observed behavior of real world networks. An example is the large number of models that have been devised that are able to produce the desired power law or lognormal degree distribution. In some kinds of networks (particularly the world wide web and social networks within corporations) there is a pronounced hierarchical structure that is not predicted by most simple models. In this talk I will describe the observed structure as some models that are able to accurately model both the degree distribution and hierarchical structure.

-----

SPEAKER: Lea Popovic

TITLE: Stochastic Models for Intra-cellular Networks

With the completion of numerous genome projects for bacteria, yeast, and humans, there is an increasing interest in understanding how molecules encoded within the genomes interact to define various functional networks of the cell. These intra-cellular networks may include gene regulatory networks, protein interaction networks, and metabolic networks.

Such networks of integrated molecular reactions tend to involve many different molecular species, where some species are present in much greater abundance than others, and where reaction rates between the complexes may vary over several orders of magnitude.

For practical purposes it is essential to reduce both the model and computational complexity of the problem, while still capturing all the essential characteristics and potential behaviour of the network.

With these issues in mind, we aim to develop and analyse stochastic models for such networks which can be well approximated by a multi-scale model of reduced complexity.

Specific issues that arise and need to be addressed include scaling limits based on the wide range quantitative scales in the system, model reduction through scaling limit approximations. The most challenging task is in analysing the implications of combinatorial restrictions which the reaction network places on the system.

-----

SPEAKER: Marko Puljic

TITLE: Synchrony in the Probabilistic Cellular Networks

Abstract: Synchronization of the firing of widely dispersed neurons appears to be necessary for the emergence of spatial patterns of cortical activity, giving direction to the neural populations that produce behavior. Using the neuopercolation models (special cellular automaton with additional distant connections that can be inhibitory and excitatory, and with sites that obey probabilistic rule), the synchrony of neural activation that is punctuated by episodes of de-coherence can be simulated. In addition, the model improves the understanding of the relationship between the system's stochastic components and the global emergent behavior. Usual methods from statistical physics and Fourier transform analysis are used to learn about the synchronization in the complex systems. When there are enough excitatory and inhibitory connections, the system components are capable of getting synchronized. But too high noise, influencing components' decisions, makes components more and more de-synchronized. There are noise levels for which the components are sometimes in the synchrony and sometimes out of the synchrony, even though the variables describing the system do not change.

-----

SPEAKER: Mason Porter

TITLE: A network analysis of committees in the United States House of Representatives

Abstract: Network theory provides a powerful tool for the representation and analysis of complex systems of interacting agents. Here we investigate the United States House of Representatives network of committees and subcommittees, with committees connected according to "interlocks" or common membership. Analysis of this network reveals clearly the strong links between different committees, as well as the intrinsic hierarchical structure within the House as a whole. We show that network theory, combined with the analysis of roll call votes using singular value decomposition, successfully uncovers political and organizational correlations between committees in the House without the need to incorporate other political information.

-----

SPEAKER: P. R. Kumar

TITLE: Scaling laws in information theory for wireless networks

I will address the following questions:

- (i) How much information can a wireless network transport?
- (ii) How does this scale with the number of nodes in the network?
- (iii) How should information be transported across wireless networks?

(Joint work with P. Gupta and L-L. Xie).

-----

SPEAKER: Raissa D'Souza

TITLE: Competition-Induced Preferential Attachment

Abstract: Models based on preferential attachment have had much success in reproducing the power law degree distributions which seem ubiquitous in both natural and engineered systems. Here, rather than assuming preferential attachment, we give an explanation of how it can arise from a more basic underlying mechanism of competition between opposing forces.

We introduce a family of one-dimensional geometric growth models, Constructed iteratively by locally optimizing the tradeoffs between two competing metrics.

This family admits an equivalent description as a graph process with no reference to the underlying geometry. Moreover, the resulting graph process is shown to be preferential attachment with an upper cutoff. We rigorously determine the degree distribution for the family of random graph models, showing that it obeys a power law up to a finite threshold and decays exponentially above this threshold.

-----

SPEAKER: Susan Holmes

TITLE: Multivariate Techniques for using Graph Structure and Covariates

In biological research it is often the case that part of a graph is known, and we want to infer other edges using covariate information on the nodes. Various multivariate techniques generalizing principal components to incorporate such information can help to solve this problem, through the eigenanalysis of a generalized covariance matrix. There has already been spectral analysis of graphs such as that by Kleinberg, we will show how these are generalized to incorporate more variables and produce more eigenvectors.

-----

SPEAKER: Volker Schmidt

TITLE: Fitting and Simulation of Models for Telecommunication Access Networks, by Volker Schmidt (University of Ulm, Germany)

We explore real telecommunication data describing the spatial geometrical structure of an urban region and we propose a model fitting procedure, where a given choice of different (both non-iterated and iterated) random tessellation models is considered and fitted to these real data.

This model fitting procedure is based on a comparison of distances between characteristics of sample data sets and corresponding characteristics of different tessellation models by utilizing a chosen metric. Examples of such characteristics are the mean total length of the edges or the number of vertices of the induced cells per unit area. We verify the algorithm by using simulated test data and subsequently apply the procedure to infrastructure data of Paris.

The talk is based on joint work with Catherine Gloaguen, Frank Fleischer

and Hendrik Schmidt.

-----

SPEAKER: George Varghese (UCSD)

TITLE: Streaming Algorithms for Traffic Analysis at High Speeds

For networks, the ability to discover patterns in traffic that can be used for better resource management, and to mitigate security threats. While offline analysis based on packet logs is being done, I focus here on online pattern detection at say 40 Gbps. In the measurement arena, the push for such real-time pattern detection comes from ISPs who have long since been plagued by the lack of assistance for managing their networks. In the security space, the push comes from the increasing cost of deploying perimeter security solutions; this has led some analysts to propose doing intrusion detection within the network. Besides these motivating forces, there is also a corresponding opportunity in terms of recent results in streaming algorithms, as well as the large amount of logic available in modern ASICs.

In this talk, after laying out this research agenda, I will try and go beyond generalities to provide some specific examples of such analysis. I first describe several component algorithms such as multistage filters, multiresolution bitmaps, and partial completion filters. I then show how these components can be put together to solve useful problems such as computing traffic matrices and detecting DoS attacks within the network, and automatically detecting signatures. Along the way, we suggest some simply stated but seemingly difficult analytical problems related to these algorithms that may be of independent interest. algorithms that may be of independent interest.

-----

SPEAKER: Kevin Lang (Overture)

TITLE: Cuts and Balance in Power Law Graphs

Using Spectral, SDP, and flow-based graph partitioning methods, we make scatter plots of cut quality versus cut balance in several real-world power-law graphs including old favorites like the internet and the web, and also a very large social graph from Yahoo.

We empirically confirm a theorem of Chung and Lu which says that certain power law graphs have an octopus structure, with a high-expansion core and weakly connected tentacles. We also see that between these two extremes there is a tradeoff populated by families of nested cuts.

We claim that these plots provide a richer view of a graph's cut structure than earlier plots of "clustering coefficient", which is still a rather local property. They are especially better than reporting a single value for the expansion of a graph, for the reality is that

each graph contains a wide range of different expansion values depending on the scale that one is considering.

Finally, we use these scatter plots as a new diagnostic tool to see whether various popular models of power law graphs replicate the cut structure of real graphs.

-----

SPEAKER: Mark Newman (University of Michigan)

TITLE: Spatial networks

Many networks exist in real space, typically on the two-dimensional surface of the Earth, and the geographical positions of network nodes can have a substantial impact on the topology of the network. As a result, the mathematics of networks in geographical space can be quite different from that of non-spatial networks. This talk will describe some recent results concerning spatial networks, including empirical data on transportation networks, the Internet, distribution networks, and others, models of spatial networks based mostly on optimality criteria, and some new methods for representing spatial networks visually that allow, for example, for non-uniform population densities. Among other things we derive some new dimension-dependent scaling laws for spatial networks, evidence of both aggregation and anti-aggregation in particular cases, and evidence that real networks are surprisingly close to optimal in a number of important respects.

-----

SPEAKER: Rick Durrett (Cornell)

TITLE: Life in a small world

In 1998, Watts and Strogatz introduced the "small world" model which is a regular lattice modified so that sites have some long range connections. This is natural from a modelling point of view since individuals not only interact with others that live nearby but also have contacts at school or at work. In this talk we will be interested in how the presence of long distance connections changes the behavior of various processes: percolation, the Ising model, contact process, random walks, and the voter model.

-----

SPEAKER: Mike Steel (Canterbury, NZ)

TITLE: Random autocatalytic networks

Abstract: The ability of systems of molecular reactions to be both collectively autocatalytic and sustained by some ambient 'food source' of simple molecules may have been an essential step in the origin of life. The probability that such systems could arise by chance in a sufficiently complex random ensemble of molecular reactions is a theme that has been explored by a number of authors over the last two decades. In this talk I outline some recent results (in papers last year with Wim Hordijk and Elchanan Mossel) on formally modelling self-sustaining autocatalytic networks. We present some results that answer simple computational and stochastic questions concerning these networks, along with some comments on

where further analysis might be helpful.

-----  
SPEAKER: Walter Willinger (AT&T Labs-Research)  
TITLE: The Many Facets of Internet Topology

ABSTRACT:

The Internet's layered architecture gives rise to a number of different topologies, with the lower layers defining more physical and the higher layers more virtual/logical types of connectivity structures. In this talk, I will show that these topologies are very different and that successful Internet topology modeling requires annotating the nodes and edges of the corresponding graphs with information that reflects their network-intrinsic meaning. To illustrate, I will focus on the Internet's router-level topology (i.e., the physical connectivity structure of today's Internet, where nodes are routers/switches and links represent physical connections) and show that it results from very structured and highly optimized tradeoffs between real-world economic and technological objectives and constraints. These findings directly contradict many popular claims, particularly those based on "scale-free" network models that ignore all such engineering tradeoffs and instead emphasize randomness (e.g., preferential attachment) and "universality." Both approaches yield network models that exhibit power law-type degree distributions (for very different reasons, though), but are in all other respects completely opposite from one another, with important implications for studying problems such as virus/worm propagation in the Internet or routing protocol performance. I will also discuss the relevance of these findings for modeling more virtual types of Internet topologies such as AS graphs, where nodes represent entire Autonomous Systems (AS) and links reflect existing peering relationships, or overlay networks like the WWW, where nodes are web pages and edges reflect existing hyperlinks.

-----  
SPEAKER: Balaji Prabhakar (Stanford University)  
TITLE: Some engineering uses of randomization and power laws

The high operating speeds in the Internet core make it a challenging environment for designing simple, high performance algorithms. (For example, the time available to process packets in the Internet core is roughly 50 ns.) Only the simplest algorithms stand a chance of being implemented; but they cannot be too simple, or they may not perform well. This talk illustrates the use of randomization and the fact that network traffic obeys power laws for designing simple, efficient algorithms.

Specifically, we shall show that randomized algorithms help simplify the implementation because they base decisions on a small random sample of the state rather than the whole state. And power laws imply the following 80-20 rule: 80% of the work is brought by 20% of the flows. Heavy advantage could be taken of such a statistic if we could identify the packets of these dominant flows with minimal overhead.

-----

SPEAKER: Dmitri Znamenski

TITLE: Connectivity, component sizes and distances in the power law random graphs.

Abstract:

Consider an i.i.d. sequence  $D_1, \dots, D_N$  of which the tail of the distribution function is regularly varying with exponent  $\tau > 1$ . We define a random graph with  $N$  nodes such that node  $i$  has degree  $D_i$ , first by assigning  $D_i$  stubs to node  $i$ , for every  $i$ , and then connecting the stubs at random.

The talk is based on three papers written with Remco van der Hofstad and Gerard Hooghiemstra (TU Delft) on topological properties of the random graphs in the limit then  $N$  tends to infinity.

First we give sufficient conditions when the random graph is connected with probability  $1 - o(1)$ . Then, for other cases, we estimate the sizes of the connected components. Finally, we derive the distribution of the number of edges between two arbitrary nodes (also called the graph distance or the hopcount), and give estimates of the diameter.

-----

SPEAKER: Chris Wiggins

TITLE: Predicting Evolution from Topology: a Machine Learning Approach

Naturally occurring networks exhibit quantitative features revealing underlying growth mechanisms. Numerous network mechanisms have recently been proposed to reproduce specific properties such as degree distributions or clustering coefficients. We present a method for inferring the mechanism most accurately capturing a given network topology, exploiting discriminative tools from machine learning. The *Drosophila melanogaster* protein network is confidently and robustly (to noise and training data subsampling) classified as a duplication-mutation-complementation network over preferential attachment, small-world, and a duplication-mutation mechanism without complementation. Systematic classification, rather than statistical study of specific properties, provides a discriminative approach to understand the design of complex networks.

-----

SPEAKER: Shweta Bansal.

TITLE: The Spread of Infectious Disease through Contact Networks.

Abstract: Contact network models attempt to characterize every interpersonal contact that could lead to disease transmission in a community, while still keeping the model tractable. Recently, the area of contact network epidemiology has come a long way (with tools like percolation theory) in removing the assumptions of traditional epidemiological models and in increasing the complexity reflected in current ones. In this talk, we review some of the recent successes in the field, and talk about exciting challenges for the future.

-----

SPEAKER: Michael Mitzenmacher

TITLE: New Directions for Power Law Research

We propose that there are five stages in power law research: observe, signify, model, validate, and control. We argue that we are currently heavily in the modeling stage, but that the research agenda should move toward the problems of validation and control.

-----

SPEAKER: John Byers

TITLE: Unveiling Hidden Topologies: Applications, Algorithms and Measurements

We consider a set of applications in Internet topology measurement, bioinformatics and physics in which the aim is to identify statistical or structural properties of a network whose nodes are either fully or partially known in advance, but whose edge connectivity is not known. In exact versions of the problems, the goal is to minimize the number of application-specific probes needed to correctly identify a subtopology (e.g. a hidden matching in one gene sequencing application). In approximate versions of the problems, the goal is to maximize the accuracy of estimates produced on a fixed measurement budget (e.g. parameters describing the degree distribution of an Internet topology). In addition to discussing algorithms and lower bounds, we consider the role played by measurement artifacts such as sampling bias, and the impact that modeling assumptions have on algorithmic performance.

-----