CS224W Reaction Paper

Beyang Liu: beyangl

October 6, 2010

1 Summary

The three papers discussed below present very different perspectives on network analysis and modeling. The first, Kleinberg et al. 2000 [2], constructs a very clean mathematical model that accurately reflects certain macroscopic properties of real-world networks, in particular the tractability of decentralized search. The second, Leskovec et al. 2008 [3], constructs a model that captures the microscopic evolution of social networks over time and shows that this process generates networks that have macroscopic properties that are similar to those of real-world networks. The third, Adamic et al. 2005 [1], studies a particular network community in depth and highlights interesting characteristics of this network.

Though they cover very different material, these 3 papers all seek to identify underlying principles that lead to network structures observed in the wild. The first 2 present mathematical models to do so, while the third is primarily a descriptive study. The last 2 evaluate on large real-world datasets, while the first only reports macroscopic statistics that seem to reflect real-world quantities.

1.1 The Small-World Phenomenon: An Algorithmic Perspective [2]

The discussion in Kleinberg et al.2001 is motivated by the goal of constructing a formal network model that exhibits the "small-world" properties of real social networks. Previous work has produced models that exhibit both the small diameter and local structure of small-world networks in the wild. Kleinberg contends that one important property has been missing in these prior discussions. This is the algorithmic small-world property – the fact that paths of short length (polynomial in the number of nodes) not only exist in the network but are *discoverable* by a decentralized search algorithm. Kleinberg goes on to define a decentralized search algorithm as an algorithm in which the current node (the one currently holding the message) has knowledge of:

1. the local contacts of every node (i.e. the underlying local structure of the graph, which could be, for example, a grid);
2. the location of the destination (where location is with respect to some "geographic cue", e.g. in a grid, this would be the (x,y) coordinate of the destination);
3. the locations and long-range contacts of all nodes that have come in contact with the message. (Note that this last property is included in all the lower-bound proofs, but is ignored in the upper-bound proofs.)

He then goes on to prove a lower bound $\Omega(n^{2/3})$ for the expected discoverable path length in a grid version of the standard Watts-Strogatz small-world model. (The conclusions generalize to non-grid graphs, but the grid is used to make things easier to discuss.) Then, he introduces a model, parameterized by a variable r that controls the locality preference of the long-range contacts of each node. The Watts-Strogatz model is a special case of Kleinberg's model (where $r = 0$). Kleinberg shows that for a different (and, in fact, unique) value of r, there is a provable lower bound $O(\log^2(n))$ on the expected discoverable path length.

The upshot is that with a value of r that matches the dimensionality of the grid (the rate at which S_n grows over n, where S_n is the number of nodes within n hops of the source node using only local edges),

we can prove an upper bound on the expected discovered path length that is polynomial in $\log(n)$. For every other value of r, we can prove a lower bound that is exponential in $\log(n)$. Intuitively, this makes sense, because setting r to this value makes it such that the probability that a long-range edge will point to a node that is n local hops away is uniform across all n. One can describe this phenomenon in the following way: When long-range contacts are long enough to offer good movement through the network but not so long that we cannot take advantage of locality cues in the underlying network, then we will be able to discover very short paths (if they exist).

1.2 Microscopic Evolution of Social Networks [3]

This paper presents a model of network evolution at the microscopic, edge-by-edge scale. It then evaluates this model on 4 large-scale, real-world social networks using the Maximum Likelihood objective. (Notably, the use of MLE is tractable, because we are looking at the per-edge process, rather than macroscopic snapshots, which would require a lot of summing out.) The model is divided into three main components:

1. Node arrival and "lifetime" selection
2. Edge initiation
3. Edge destination selection

Node arrival rates are heavily dependent on the particular network, and so this is specified as input into the model. Node lifetimes describe the span of time over which a node is initiating new edges, and this is drawn for each node independently from an exponential distribution. A newly arrived node adds its first edge immediately, using Preferential Attachment to choose its new neighbor (this means that it prefers to attach to higher-degree nodes, but otherwise chooses at random from the graph). The node then "sleeps" for a period of time drawn from a power-law-with-exponential-cutoff distribution. When the node wakes up, if it is still within its lifetime, it initiates a new edge. This time, it chooses its neighbor using a local triangle-closing process and then goes to sleep again.

Of particular interest is the mechanism by which a node initiating a new edge chooses its new neighbor. Notably, all new edges close two-hop gaps (i.e. they create new triangles). This restriction is not unrealistic, because the data exhibits exponential fall-off in the number of new edges that close n-hop gaps. In addition to choosing a node two-hops away uniformly at random, the paper discusses a few more sophisticated algorithms for choosing new neighbors. All of these involve first choosing a neighbor and then having that neighbor choose one of its neighbors. Evaluating against the log-likelihood of edge formation in the data, the paper concludes that algorithms incorporating degree information and recent node activity perform best, but a simple model in which neighbors are selected uniformly at random in both stages performs comparably.

The upshot is that local preference alone is a strong guiding principle of network evolution, and that preferential attachment and other forms of preference play, at best, secondary roles.

The paper concludes by showing that the microscopic process of its model produces networks that exhibit macroscopic properties (clustering coefficient, degree distribution, geodesic distance) that more accurately reflect real-world networks than previous models (in particular, models of Preferential Attachment). Finally, it notes that its model can be used to generate realistic networks of arbitrary size.

1.3 The Political Blogosphere and the 2004 U.S. Election [1]

This paper explores the network properties of the political blogging community in the months leading up to and following the 2004 U.S. election. The main observation is that the subnetworks of liberal and conservative blogs are segregated along ideological lines and that the conservative subnetwork is more densely connected (but not overwhelmingly so) than the liberal one.

A good portion of the paper describes the data collection process. The data is divided into a large coarse-grain snapshot of hundreds of blogs and a fine-grain per-post dataset on a subset of 20 liberal and 20 conservative "A-list" blogs.

The following is a list of key observations about the data:

- In the top 40 blogs, cross-linking between liberal and conservative sites was rare, while intra-linking was higher within the conservative network.
- Though conservative bloggers tend to refer to each other more often, the uniformity of what they say (measured by a cosine similarity of URLs mentioned in posts) is slightly lower that of liberal blogs.
- Given a political figure (e.g. George Bush), it is more likely that there are more mentions of him/her by bloggers of the opposite political orientation, suggesting that bloggers find it easier to criticize their opponents' policies rather than evangelize the policies of their own policymakers.
- Statistics on the number of links from blogs to mainstream sources illustrate that liberals and conservative bloggers choose to reference a different set of mainstream sources.

2 Critique

2.1 The Small-World Phenomenon: An Algorithmic Perspective [2]

Kleinberg presents an excellent generic model that is capable of capturing the small diameter (i.e. existence of short paths), local structure, and searchability (i.e. discoverability of short paths) of real-world social networks. The model he presents is very clean and simple, but it also generalizes to other categories of networks (in addition to grids, the paper also mentions trees). Despite this mathematical generalizability, however, it seems that the class of networks that fall into the Kleinberg model lack some richness present in real-world networks, and by including this richness, we could create a better model that more accurately reflects reality. For example, Kleinberg mentions in the Related Work section that previous work has studied the passage of a message through different social "categories." In the real world, some individuals are more well-connected than others (this could be due to the quality of their personalities, their socioeconomic standing, or some official hierarchy built into the network). The connectivity of a node provides another heuristic (in addition to Kleinberg's "geographic" cues) that could be used in a decentralized search algorithm. Furthermore, studies of the Preferential Attachment phenomenon suggest that a broad distribution of node degrees is a property that naturally arises from the evolutionary process of networks [3]. In the Kleinberg model, all nodes have the same degree. While the spirit of the paper's conclusions probably generalizes to graphs with a different degree distribution, it is unclear that the specific bounds would. A possible next step is to construct a model that exhibits all the preferable properties of the Kleinberg model while simultaneously producing a realistic degree distribution.

2.2 Microscopic Evolution of Social Networks [3]

This is the first study of social network evolution on a microscopic (i.e. edge-by-edge) scale and the fact that the model presented produces both microscopic and macrosopic properties present in real-world social networks makes this a seminal work. The paper evaluates against several important macroscopic metrics, such as clustering coefficient, degree distribution, and geodesic distance (but not, however, decentralized searchability). In addition, the novel maximum likelihood objective over each edge-arrival offers a new standard metric for evaluating temporal, graph-evolution models.

Though the model works exceptionally well in the case of the 4 online social networks presented, there is no reason that its fundamental components could not be generalized to encompass other types of networks. These fundamentals are the idea that graph evolution can and should be analyzed on a microscopic scale (instead of a series of macroscopic snapshots) and that the process can be subdivided into 3 sub-processes: node arrival, edge initiation, and edge destination selection.

In particular, other types of networks to consider include older networks. These networks, unlike the 4 presented, may have low or even negative growth. Thus, there would have to be some mechanism to model node death (which is explicitly ignored in the paper). Notably, the Leskovec model seems to perform most poorly on the Answers network, which saw a lower growth rate than the other three, which suggests that there is room for improvement in the model's generalizability.

An example of an online network that might not follow the explosive growth of the typical social network is the network of political bloggers, who have been around for awhile and whose network is probably more static. Such older networks are probably more prevalent in the physical world (rather than in online communities), and though most probably do not offer the fine-grain temporal datasets necessary to evaluate this type of model, there are a few that might. For example, the FEC collects data on the flow of donations from lobbying groups and individuals to political candidates. The granularity of the data is at a per-transaction level, and it might be interesting to study the evolution of the network of interest groups and policymakers over time.

2.3 The Political Blogosphere and the 2004 U.S. Election [1]

The paper presents an overview of interesting statistical properties of an extensive dataset of the political blogging community, but does very little beyond preliminary commentary on these properties. There is no attempt to model the interactions or make predictions about future events. Due to the lack of a big picture analysis, the strength of the arguments made seems tenuous and vague. For example, the paper comments that the conservative subcommunity seems more densely connected (but not overwhelmingly so) than the liberal subcommunity. Little insight, however, is provided to connect this structural network property back to physical reality. Large questions loom and remain unanswered. How did this difference arise? Is it the result of a conscious effort or is it, perhaps, something that arises from the ideological dispositions of each faction? What are the effects of this denser connectivity?

Furthermore, though there is some analysis of blog topics and how they relate to current events (e.g. the CBS Bush war documents debacle), it would be more interesting if the temporal data covered a broader timespan. In particular, the reader is left to wonder how the blogging landscape has changed since the publishing of the paper. What was the blogging landscape like in 2008, when election results were far different? How are they now as the new administration struggles to implement its new policies? What would be even more interesting and useful would be to try to study the interaction of political victories and activity on the blogosphere. Does blogging intensity precipitate political victories, or vice versa? How can we quantify, and ultimately model, the interaction between actual election outcomes and the informal political press?

It would also have been nice to utilize the fine granularity of the blog post data to study how political events unfolded, at a very fine scale, across the online political world. Taking inspiration from the Leskovec paper, it would have been nice to try to model the structural evolution of the blogging community at a microscopic level, as well. In particular, it might have been interesting to try to uncover correlations between cross-referencing frequency and the growth of most-cited blogs.

3 Brainstorming

One of the appealing qualities of the Leskovec paper is its marriage of theory and practice. It constructs a precise and elegant model that actually reflects network properties observed on four very large datasets. Along these lines, it would be nice to apply the theoretical ideas of the Kleinberg and Leskovec papers toward a hitherto untapped domain.

One such domain is that of political and policy analysis. Previous work has looked at the intersection of the Internet and politics, but networks that fall into this area (e.g. blogging communities) are at best secondary indicators of the political situation.

Very few studies have aimed at quantitatively studying the physical social networks of policymakers and lobbyists (and in particular, the cash flows upon which such networks are founded), and of the few studies that have done so, many have been from the perspective of political or social science. To my knowledge, there has not yet been a large-scale objective analysis of data on this topic (even though such data – at least the legal and official portion of it – is publicly available).

There seems to be a general consensus that the network of policymakers and lobbyists at the federal level has grown in size and complexity over the past decade or two. It would be nice to confirm this anecdotal evidence with statistics. Furthermore, it would be interesting to study the structural properties of such a network. Do 2 distinct subcommunities arise? How many cash-flow hops away is any given

lobbyist from any given candidate (network diameter)? What is the clustering coefficient? What is the node degree distribution? How does the network evolve over time (both on a macro and micro scale)? Can we predict new cash flows in the future?

Finally, given some model for such a network, it would be interesting to apply machine learning techniques to predict policymaker behavior. Knowing that certain lobbying groups support certain bills, can we predict the probability that a certain bill will pass?

In our modeling task, we could either take advantage of prior structural knowledge of our networks (e.g. the fact that there are 2 major political parties) or ignore such prior knowledge and attempt to "discover" it through generic network analysis.

References

[1] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *In LinkKDD 05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *in Proceedings of the 32nd ACM Symposium on Theory of Computing*, pages 163–170, 2000.

[3] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.