Viewing Implicit Social Networks As Bipartite Graphs

Benjamin Bercovitz
Stanford University
CS322
berco@cs.stanford.edu

ABSTRACT

In this paper, we describe a new model for thinking about implicit social networks, which are defined when users interact by modifying a common shared object. These are also called "knowledge sharing" networks and are in contrast to the typical person-to-person social networks. The properties of these networks are important to many data mining applications including recommendation systems, personalization, and other networks. The model consists of a bipartite graph where users and interaction points are two classes of nodes, connected by edges when the users interact with the shared objects represented by the interaction points. For example, an interaction point for an online book store might be the reviews for a certain book. In the end, we conclude that the information in the network structure of the bipartite graph contains significant useful information on the implicit social network and can be used as an effective data structure for analysis and visualization.

1. INTRODUCTION

Social networks are usually studied by modeling them as graphs, where nodes represent people and (possibly weighted) edges between the nodes represent the connections between people. This view is generally useful in the study of social networks because the primary objects that we want to study are represented by nodes in the network. Therefore, analyzing the graph properties informs some understanding of the properties of human networks and behavior.

Not all social networks, however, consist of direct relationships between people. Networks formed by Facebook, LinkedIn, email, etc. are formed by "active" interactions; that is, an interaction is passed directly from one user to the other with the sender's intent being to raise the attention of the receiving party. In contrast, other social networks are formed by passive interactions, or interactions that are propagated by the common (but possibly or even likely independent) accessing or modifying of a shared object or datum, which defines an *interaction point*. Web site users might never know each other personally but nonetheless form a implicit network through shared interactions. For instance, a web site that lets users post book reviews would gradually collect an implicit social network when a pair of users review the same books.

Arguably these are actually the most common type of social networks on web sites today. One key aspect of these services is that they are often anonymous (or pseduononymous, where users have a pseudonym or screen name that is consistently used to refer to them), allowing users to more freely share their opinions or ideas. Web site owners may prefer such services because they think it produces more valuable content for their web site. They may also prefer it because it limits the ability of groups to organize and protest against products or policies that they see as unfair, hurting the reputation of the company. The other key aspects of implicit social networks are that they form by different users acting on the same shared object or data and that there need be no prior relationship for this interaction to occur.

Indeed, these networks are not just prevalent, but they are also valuable. The information contained in these networks is commonly used for personalization, predictions, recommendations, and clustering. Product ratings, click graphs, textual reviews, and wish lists are all interaction points used to compare one user's actions to other similar users in order to extract some likely generalizations that can be applied to the original user. The view this paper takes is that these can best be understood as a bipartite network with two classes of nodes: users and interaction points. The edges between these node classes are mapped one-to-one with interactions that take place.

This view contrasts with two prevailing views on analyzing user data. In the data mining view, changes to shared objects are represented as attributes or feature vectors owned by the user objects. The review of a book is filed under the user who wrote it, essentially. Then, using concepts such as frequent itemset analysis, or other set-based or probability based techniques, this information is processed to obtain likely predictions. The disadvantage of this technique is that it might not take a network-like view of the users and misses out on potential connections and properties that would have been present in a network model.

Also in contrast with the proposed view is the standard social network view, where users are nodes and their connections are edges. To analyzed users and gain knowledge about them, typically their interactions are mined and reduced to edge weights (or, perhaps, multidimensional edge weights). Effectively, the background structure that connects two user nodes is collapsed into a single number, such as a similarity measure.

The goal of this work is to explore what information is contained in the structure of the proposed bipartite model for implicit social networks and find out what, if any, network models are useful in understanding their properties. There is an expectation that user nodes will tend to behave as they do in the standard social network model, which means that they have similar properties as a power-law random graph at the global level, but have significant local structure like a small world model and specialized network evolution properties like a preferential attachment model. As for the interaction point nodes, it is unclear what assumptions can be made about their connections to the individual user nodes.

2. RELATED WORK

The model of Onuma, Tong, and Faloutsos [1] on TANGENT was quite intriguing. The main result was a new method for generating "surprising but relevant" recommendations by looking for bridge nodes between groups that form around certain user preferences. The anecdotal example lend themselves to understanding the intuition behind the idea, but the reader is left without being convinced that a nearby bridge node is actually a good recommendation. Despite this shortcoming, the idea of treating the MovieLens social network as a bipartite network makes perfect sense and this viewpoint might be applied to other networks as well.

Guo, *et al.* [2] works on understanding the generation of user-provided content on knowledge sharing networks in light of recent models like the one presented in Leskovec, *et al.* [3]. Its main results are two. First, the user lifetime does not fit an exponential distribution like in "network oriented" social networks studied by Leskovec [3]. Second, the amount of content created per user follows a stretched exponential distribution overall that approximates a power law when the parameter *c* is small. The authors conjecture that *c* might be related to quality or effort required to share the knowledge. Also of note, the "core" users have a different profile from the majority of users. Whereas the majority's contribution drops off exponentially quickly as the number of contributions increases, the core users contributions fit a flatter profile more like a power law. Overall makes a valuable insight that "networking oriented" online social networks might be formed by different forces than "knowledge sharing" online social networks and have a different behavior overall.

3. MODELS

The proposed model is called the bipartite graph model for implicit social networks. Implicit social networks are defined by interactions that users make with shared data objects ("knowledge sharing") called *interaction points*. The users do not interact directly with each other (though they will passively know that others are interacting with the same objects; this is how they develop structure). Obviously, the interaction points do not interact with each other.

This interpretation lends itself to the obvious representation as a bipartite graph, with two classes of nodes. An inspirational example of the proposed model is shown in Figure 1. On one side, the nodes represent individual users. On the other side, nodes represent the interaction points, such as books reviewed. The presence of an edge from a user to an interaction point means that the user has experienced an interaction with that shared object. Metadata about the interaction can be added to the edge, such as the number of stars. Note how this is a lossless interpretation of the original interaction.

Network properties can now be measured over either set of nodes to paint a clear picture of the data, or the entire network can be projected on to a single class of nodes and analyzed as a typical social network graph would be (while losing any metadata if it cannot be systematically combined).

Computing global network properties is not a problem, but determining local structure is a problem because bipartite graphs have no local structure. Furthermore, the neighbors are not of the same node type. If a single node class projected view is taken, then we can investigate "local" structure, but it will not be really local because a node that looks two hops away is really four hops away. This is a material difference in terms of network models. A good compromise is to use a *bipartite clustering coefficient*.

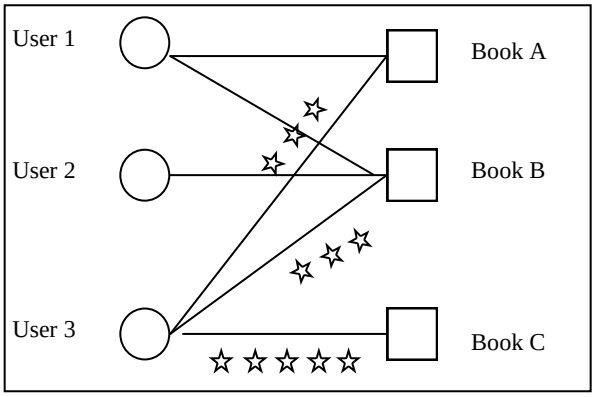

Figure 1. Example of the bipartite graph model

Consider neighbors of User 3 in Figure 1. The standard clustering coefficient is the fraction of possible edges that exist between the neighbor set. Clearly this is zero in this example. So we will try the next-best thing, first making a hop back to the users. In this example, (A,B) would be counted as a closed triangle, but not (A,C) or (B,C).

4. EXPERIMENTS

There are three different kinds of networks analyzed here and summarized in Table 1.

The first type is item tagging networks. Data is available from Flickr [5], Delicious [5], and Last.fm [4]. The interaction points are items, with the tags used as edge metadata.

The second type is item rating networks. Data is used from the Netflix Prize dataset [7] for this network type. The interaction points are movies, with the ratings as edge metadata.

The third type is citation networks. Data is used here is from a high energy physics group [6]. The user nodes are citers (papers that cite another) and the interaction point nodes are citees (papers that are cited). A paper could wind up with a node in both if it both cites and is cited.

The experiments were broken down into three parts. First, each graph was analyzed for global structures. Degree distributions were analyzed, as was the largest connected component size versus number of remaining nodes and compared with a rewired version.

Next, selected graphs were analyzed for local structure using plots of the bipartite clustering coefficient over different numbers of remaining nodes and compared to a rewired network.

Finally, a selection of graphs were re-examined with edge arrival information to see if any conclusions could be made about generative models. Since the lifetime is infinite in a sense, and the arrival rate is well-studied and not affected by the bipartite view, the main focus here is on the destination selection for new edges.

Table 1. Network Types in Experiments

Node Type	Tagging	Movie Rating	Citations
User	User	User	Citer
Interaction Pt	Item	Movie	Citee

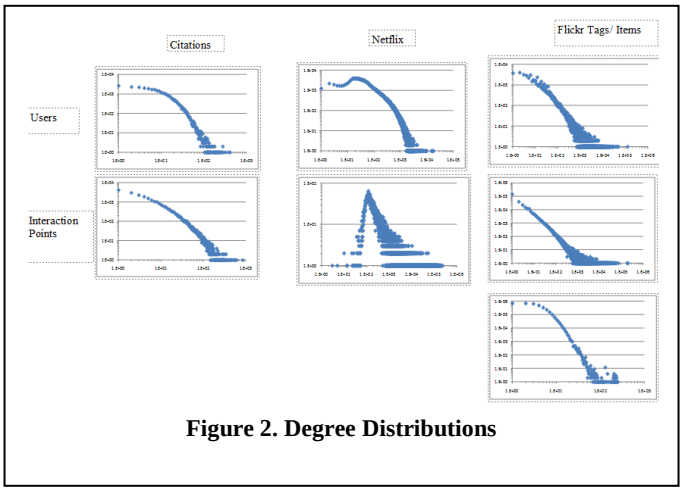

Figure 2. Degree Distributions

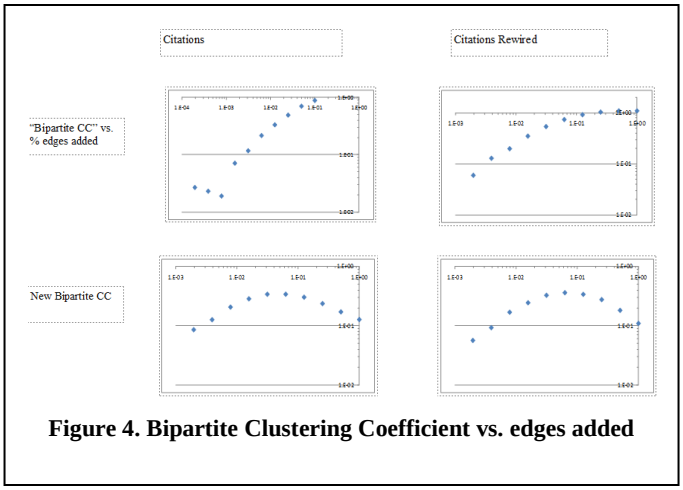

Figure 4. Bipartite Clustering Coefficient vs. edges added

5. RESULTS

In each data set, there were differing ratios of nodes on each side of the bipartite graph. For the movie rating network, there were many more users than movies. For the item tagging network, there were many more items than users. Finally, for the citations network, there were nearly equal numbers. This diversity allowed the exploration of the different ratios.

To analyze the networks, a new software library was created to facilitate the gathering of statistics on bipartite networks. One of the stumbling blocks was the sheer size of the networks, which precluded fully loading them into memory even on a fairly good (6GB RAM, 4CPU) machine. The network nodes were stripped down to bare arrays and integers to fit the whole networks in memory. The software library was able to perform sorting using a multi-pass approach where intermediate blocks were written to disk (much the same as the well known sort program in UNIX). Even so, some of the computations were too expensive on the larger graphs.

5.1 Degree distribution

The first evaluation step was to look at the degree distributions of the various graphs and subgraphs, which are shown in Figure 2. Overall, all of the data sets exhibited a power-law distribution at least in the tail. Various data sets clearly have a cutoff on the lower degree counts, and the Netflix data set has a particularly strong cutoff. This is obviously due to the fact that movie rental company cannot afford to rent movies that fewer than a few

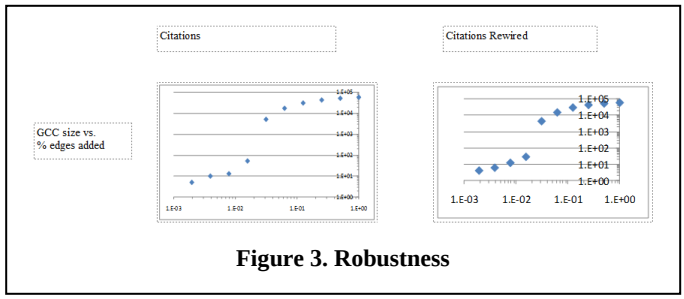

Figure 3. Robustness

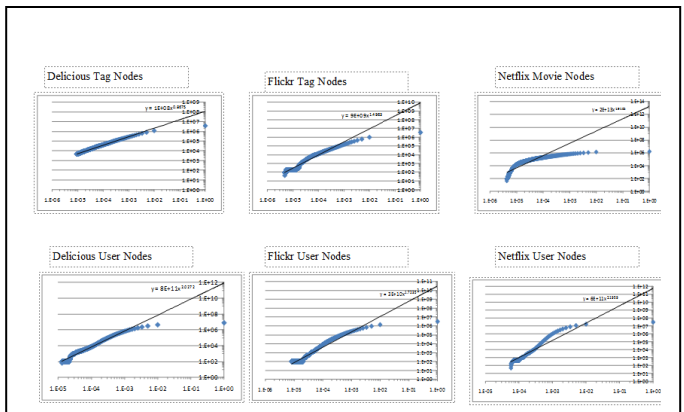

Figure 5. Network Evolution: Probability vs. Dest. Degree

hundred people will watch (and that they are particularly good about collecting reviews).

These data are mostly consistent with the view that from the perspective of both classes of nodes, the networks have a global structure consistent with a scale-free random graph. This raises the question, is it necessarily true that if one class of nodes follows a power law, that the other must also? It can be shown that the distribution of one class of nodes does not imply the distribution of the other, so long as the sum of the out degree matches. It is entirely possible to have one class of nodes have a power law degree distribution and the other class have constant degree or a binomial distribution like the Erdos-Renyi model. This did not happen in practice in any of our data sets, however.

5.2 Robustness and local structure

The next evaluation was a robustness evaluation. Nodes were gradually removed until none remained. At each step, the graphs was split into connected components. The size of the largest connected component is plotted against % of nodes remaining in Figure 3. The evaluation was only done on the citations network because the size of the larger networks prevented computational feasibility. The result is consistent with a rewired version of the same graph, so that indicates that this property was not the result of any local structure, but is rather a global property.

The bipartite clustering coefficient was the next to be measured. As with robustness, it was measured for varying ratios of remaining nodes, and also compared against a rewired version of the same graph. Both plots can be seen in Figure 4. Once again, the evaluation was only done against the citations network because of the size involved. This allowed the isolation of local graph structure. Both the trend and the relative values of the bipartite clustering coefficient indicated that either there was no

local densification or that the coefficient did not really have the resolution to measure it.

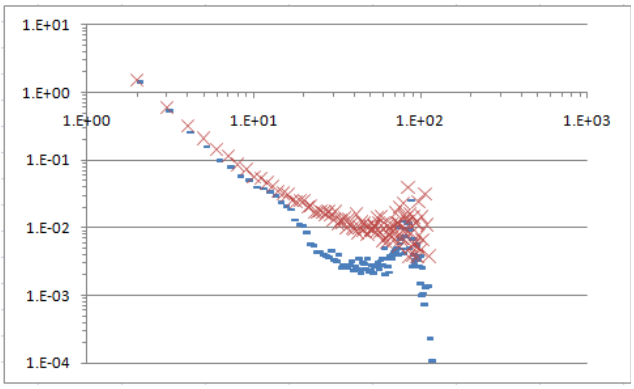

Figure 6. BCC Distribution (- mark indicates rewired)

5.3 Network Evolution

The final set of evaluations were network evolution plots. Each graph that had edge arrival information was sorted by edge arrival time and then the graph was built from scratch by adding edges in the order of their original arrival. As this was occurring, statistics were taken on the degree of the destination node. The results are shown in Figure 5. The different networks show a power law relationship between the destination node degree and the probability that the edge is added.

In general these are consistent with a preferential attachment-like view of network evolution. That would explain the distribution of the destination node probabilities. In addition, other network evolution properties, the node lifetime and arrival rate, would imply the power law distribution of node degrees if they had exponential distributions.

5.4 Local structure reconsidered

Looking at the local structure of these graphs, there does not seem to be a whole lot of it. Most of the associations between users are weak, or by chance. However, the intuitive knowledge of these networks and the fact that recommendation systems have been successfully implemented on most of them points to the fact that local structure is important. Another approach that could give better results would be to have a clustering coefficient that is normalized by the degree of the interaction point. This would make a lot of sense because it would discount the effect of very high degree edges and boost the low degree ones. Especially when combined with the network evolution results, this makes a lot of sense. Since the probability of adding an edge to a low degree node is empirically found to be quite low, then we can conclude that two users who consistently choose common low degree nodes are indeed much tighter neighbors than those connected through common high degree nodes. The re-implemented version that takes into account these weights is shown in Figure 6 and displays the overall distribution of the coefficients with respect to node degree. For comparison, the bottom half of Figure 3 is the new discounted measure, while the top half is the original measure.

The data in Figure 6 show that for low degrees, there is little local structure compared to the baseline rewired graph, but for the rare high degrees, there is indeed local structure, which confirms our expectations. This implies that taking the view of a bipartite graph for implicit social networks preserves one of the key properties used in data mining systems. Thus, this can be used as a either a visualization tool or a data structure for mining additional information.

6. CONCLUSION

Taking the view of a bipartite graphs with two node classes can help visualize and analyze data in implicit social networks. The results showed that most of the network properties are shared between the two sides of the graph, even though they need not necessarily do so. Furthermore, this paper has introduced a bipartite clustering coefficient and refined it so that it can be useful to for measuring locality in the network, even though empirically that locality was not strong for the low degree nodes that dominate the networks.

7. ACKNOWLEDGMENTS

Thanks to plamere's blog at Sun Research Labs for the Last.fm dataset [4], the PINTS project at University Koblenz-Landau for the tagging datasets [5], Stanford SNAP [6] for the citations network data and the Netflix Prize competition for the Netflix data [7].

Thanks to the CS322 teaching staff and Prof. Leskovec for a great course. Very informative and quite thought-provoking. It was the highlight of my Tuesdays and Thursdays this quarter.

8. REFERENCES

[1] K. Onuma, H. Tong, C. Faloutsos. TANGENT: A Novel, "Surprise-me", Recommendation Algorithm. In *KDD* '09.

[2] L. Guo, E. Tan, S. Chen, X. Zhang, Y. Zhao. Analyzing Patterns of User Content Generation in Online Social Networks. In *KDD* '09.

[3] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins. Microscopic Evolution of Social Networks. In *KDD '08*.

[4] plamere's blog, Sun Labs, Last.fm data from audioscrobbler http://blogs.sun.com/plamere/entry/open_research_the_data_lastfm

[5] PINTS Experiment Data Sets, Universitat Koblenz-Landau, http://www.uni-koblenz-landau.de/koblenz/fb4/institute/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_html

[6] Stanford SNAP Graph Library Data Sets http://snap.stanford.edu/data/index.html

[7] Netflix Prize Dataset, UCI Machine Learning Archive http://archive.ics.uci.edu/ml/machine-learning-databases/netflix/