DETECTING HIDDEN STRUCTURE IN SOCIAL NETWORKS

DIANA CAI AND TONY FENG

Abstract. We consider the problem of predicting highest-weight edges occurring in social networks, where the weights measure relationship strength in some way, using only inherent topological features of the graph. We apply topological metrics from the information retrieval and link prediction literature to rank edges, and compare the rankings thus obtained to the true rankings by weight. We find that for scientific collaboration graphs and Facebook, this method offers a significant improvement over random guessing when attempting to identify a small proportion of top edges.

1. Introduction

Graphs are useful objects for studying the structure of networks, but in many cases they fail to capture finer qualities of the relationships being described. For instance, online social networks represented from graphs differ significantly from real social networks: on Facebook, a user may have several hundred friends, but in real life, one seldom interacts with more than a few dozen people regularly.

In general, graphs representing interactions within a large system of objects may fail to describe interesting additional information about these interactions. We are interested in studying to what extend this hidden data can be recovered from the structure of the graphs. Our problem may be posed generally as follows. We model a network of interacting objects by a vertex set corresponding to the objects, with links weighted in a way that reflects the strength of the relationship. Let the associated unweighted graph be obtained by forgetting the weightings of the edges. For what kinds of networks and interactions can we recover the highest-weight edges from the purely *topological* data of the unweighted graph?

We focus on the specific problem of social networks. Working under the assumption that there are a few important relationships among the many edges, we analyze topological properties of the graph in attempt to deduce these import edges. Our main tools are algorithms drawn from the literature on link prediction and information retrieval that estimate the "similarity" of two nodes based on their connective neighborhoods.

For instance, one would expect close friends to have many common friends, because people who spend much time together must have similar friend circles, simply by limitations on time. Therefore, one would expect users that have many common friends to have a stronger relationship.

There are many ways of refining this heuristic. The number of edges that a particular node has speaks to its "gregariousness." Perhaps this means a greater predilection towards socializing and developing strong bonds. But one could equally imagine such a person having a tendency towards shallower relationships, spending less time on many

more connections. Different intuitive models of the problem lead to different ways of gauging similarity. Which is most effective in social networks? This is the sort of questions that we try to answer.

We ultimately find that these techniques are very effective in identifying the top strata of high weight edges in collaboration graphs of physics authors and in Facebook. When trying to find the top $2\% - 5\%$ of edges, our methods perform several times greater than blind selection. When we are simply trying to find the top 50% of edges, however, we cannot do much better than randomness. This result suggests that, for these networks, there is only structure near the top. We find that the *resource allocation* and *highest common neighbors* heuristics perform best for collaboration networks, while *resource allocation* and *Jaccard's coefficient* perform best for Facebook.

We also test our methods on an online dating site called Libimseti and an online community called Advogato, and we find that we cannot do any better than random guessing. Based on these results, we hypothesize that the heuristic of neighborhood similarity applies only to networks that have some physical grounding.

Applications. Information is power. Anybody who better understands the relationships in social networks is better poised to act in way tailored to this extra knowledge. Facebook could be more accurate in recommending groups, events, or advertisements if it better understand the dynamic between users, and the same considerations apply for any social networking site.

Another application of the ability to predict link strength from topology is to compression. If only is only concerned with "important edges," then our findings suggest that a certan amount of information about weighted graphs can be compressed away into topogical structure.

Even more significantly, there is the possibility of applying the results to graphs where no weighting is known, especially graphs that are not even social networks. For instance, one could apply these analyses to the network of connections between neurons, and inquire if the most "important" links may be uncovered in this way. Or one could study a function of multiple variables, and attempt to induce the strongest causal or correlational relationships. In many real applications, graphs capture only a rough picture of the structure of relationships and interactions, and it is desirable to recover the hidden structure.

The rest of this paper is organized as follows: in Section 2, we discuss how our work fits in with related work in predicting link strength. In Sections 3 and 4, we discuss our methodology and setup, detailing our topological methods and the metrics for which are measuring. We discuss our results from applying these methods to the real-world network data we gathered in Sections 5 and 6. Lastly, we summarize our results and discuss future directions in Section 7.

2. Related Work

The idea of predicting link strength is not a new one. Kahanda and Neville [1] study the problem of using extraneous "transactional" information, such as wall posts and

picture tagging, to predict link strength in Facebook. Xiang et al. [7] bring the additional information of user profile similarity to attack this question. Both approach the problem with machine-learning: the former employs supervised learning through logistic regression, bagged decision trees, and naive Bayesian classifiers; and the latter uses unsupervised learning of a latent-variable model. They are able to achieve impressive success, but in utilizing such specific features of the graphs that they consider, limit their results to a very narrow class of networks sharing Facebook's features. Our motivation in considering topological characterics only is to have a broader relevance.

Our work connects this line of questioning to research in the *link prediction problem*. In this framework, we consider a graph, perhaps depicting ties in a social network, and attempt to predict the next links to form. Liben-Nowell and Kleinberg [2] defined this problem, and approach it by ranking edges by *topological* metrics such as the number of common neighbors or the graph distance. They test it for *co-authorship* graphs of physicists. We adopt some of their methods for our setting. The survey of Lu and Zhou [3] gives a more comprehension discussion of the methods used in link prediction.

3. Setup

We consider a *weighted* graph G, which may or may not be directed. For a vertex $v \in G$, the *true set* $T_k(v)$ of size k consists of the k edges of v that have the highest weight.

We are interested in some topological metrics, intrinsic to the graph structure, that can be used to detect the true set in the following way. A metric gives a ranking the edges of v, producing for any k a *guess set* $G_k(v)$ of the k highest ranked edges. We define the *success rate* to be

$$\frac{|G_k(v) \cap T_k(v)|}{k}.$$

In the examples we consider, G corresponds to a social network, with the vertices representing users and the edges representing "friendships." The weights encode some intrinsic measure of relationship strength. We would like to maximize our success rate.

To put the results in perspective, we define the *random success rate* to be the success rate of the metric that simply chooses a random ordering of edges. We will then consider the ratio of the success rate and the random success rate to give a quantitative measure of how much our metric improves over random guessing.

4. Topological Methods

We describe here several metrics for measuring edge strength, which are used by Liben-Nowell and Kleinberg [2] in the context of link prediction. One basic heuristic is that close friends tend to have similar friend neighborhoods by virtue of their interactions. Therefore, metrics that are based on comparing neighborhoods of vertices are natural choices. For a vertex $v \in G$, let $\Gamma(v)$ denote the neighbors of v.

(1) **Common neighbors.** The metric is $m(\overline{xy}) = |\Gamma(x) \cap \Gamma(y)|$. Intuitively, friends introduce each other to other friends.

(2) **Jaccard's coefficient.** The metric is

$$m(\overline{xy}) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

This may be viewed as a normalized version of the common neighbors metric. More generally, Jaccard's coefficient is a similarity metric used in information retrieval to measure the probability that x and y share a feature that at least one of them has. In this case, we are considering the feature of a shared friend.

(3) **Adamic/Adar coefficient.** The metric is

$$m(\overline{xy}) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}.$$

Originally, a metric was introducd by Adamic and Adar to measure similarity of homepages by

$$\sum_{z \text{ feature shared by } x,y} \frac{1}{\log(|\text{frequency}(z)|)}.$$

Here again, we take features to be common neighbors and their frequency to be measured by the neighborhood size. of resource received by y relative to x is $m(\overline{xy})$ above.

(4) **Preferential attachment.** The metric is $m(\overline{xy}) = |\Gamma(x)||\Gamma(y)|$. The preferential attachment model posits that new edges form with probability proportion to this value, forming highly clustered graphs with a few nodes of large degree. This metric was proposed based on empirical analysis of the co-authorship graph.

(5) **Propagation coefficient/Resource allocation.** We devised a "Propagation metric" given by

$$m(\overline{xy}) = \frac{1}{|\Gamma(x)|} + \frac{1}{|\Gamma(x)|} \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|}.$$

This is motivated by a model of resource allocation between nodes. The node x is viewed as sending some resource to all of its friends, which has a secondary effect to all of the friends of the node that receive it: each of these friends gets a uniform piece. Therefore, if y receives $\frac{1}{|\Gamma(x)|}$ directly and then $\frac{1}{|\Gamma(x)||\Gamma(z)|}$ from each friend z.

We originally intended to use this in a way analogous to belief propagation in error-correcting codes, wherein nodes sent each other propagation coefficients for several rounds, and constructed new propagation coefficients for each round. The updated propagation coefficients would, more generally, be

$$m(\overline{xy}) = \frac{w(\overline{xy})}{w(\overline{x})} + \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(\overline{xz})}{w(\overline{x})} \frac{w(\overline{zy})}{w(\overline{z})},$$

where $w(\overline{x})$ is the weighted degree of x and $w(\overline{xy})$ is the weight of the edge $\overline{xy}$. The idea is that this process models the give-and-take involved in maintaining

real relationships. However, it turned out that propagating only once was superior, and the accuracy actually deteriorated with each propagation. We also discovered later that this is already a known metric, under the name of Resource Allocation.

5. Results

We tested our methods on two graphs obtained from real datasets, obtained by the following procedure. For each node x (corresponding to a user or author), we computed a number $k(x)$, and chose the $k(x)$ highest ranked edges among each metric:

(1) Random selection,
(2) Highest common neighbors,
(3) Highest Jaccard coefficient,
(4) Highest preferential attachment.
(5) Highest propagation/resource allocation coefficient.
(6) Highest aggregate.

We then compared the guesses against the $k(x)$ friends that had the true highest edge weights, and computed the fraction of the guesses that were correct in each case.

The numbers $k(x)$ were found in two different scenarios: one where $k(x)$ is simply a proportion p of the degree of x, and another where $k(x)$ was the number of edges with weight above a certain cutoff. The second scheme performed better, since attempting to find more "strong" edges than there actually are is naturally unfruitful.

5.1. **Advogato graph.** Advogato is an online community directed towards the development of free software [4]. Relationships are rated by users on three levels based on trust level: Apprentice, Journeyer, and Master. We map these to edge ratings $1, 2, 3$, respectively.

For each user x in the network, we found the number $k(x)$ of other users with the maximal trust relationship (Master), and computed $k(x)$ random guesses from all of x's neighbors in addition to the top $k(x)$ according to each metric. We then compared these against the true set of Masters. None of the metrics exhibits any significant improvement over random selection. In Table 1 we exhibit the ratio of correct guesses over the whole network to the number of correct random guesses.

Common	Jaccard	Adamic/Adar	Preferential	Aggregate
0.965	0.957	0.971	1.043	0.967

Table 1. Factor of high weight links identified in the Advogato network over random selection.

5.2. **Libimseti.** Libimseti is an online Czech dating site in which users can rate others on a scale from 1 to 10. As with Advogato, we found that none of our metrics worked better than random selection, on average.

5.3. **Coauthorship Graphs.** We have three graphs representing coauthorships in astrophysics, condensed matter, and high-energy theory collaborations from 1995 to 1999 [5]. The edges represent collaboration, and are weighted by the number of collaborations in the following way: if authors x and y collaborate in a paper with n authors, then the edge weight between x and y is increased by $\frac{1}{n-1}$. This is summed over all co-authored papers written in the relevant period of time. This data was obtained by crawling the arXiv.

In Tables 2 and 3, we show the results averaged over 1000 nodes with p running from .05 to .5 in steps of .05.

p	Random	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.05	0.053	0.421	0.421	0.368	0.105	0.421	0.421
0.1	0.126	0.337	0.326	0.368	0.147	0.347	0.389
0.15	0.140	0.455	0.374	0.423	0.221	0.396	0.491
0.2	0.193	0.407	0.328	0.387	0.236	0.400	0.458
0.25	0.225	0.482	0.378	0.401	0.278	0.385	0.529
0.3	0.272	0.444	0.393	0.423	0.285	0.416	0.503
0.35	0.313	0.514	0.418	0.434	0.327	0.422	0.575
0.4	0.334	0.512	0.425	0.408	0.374	0.404	0.556
0.45	0.398	0.532	0.454	0.443	0.405	0.447	0.578
0.5	0.467	0.581	0.510	0.490	0.486	0.492	0.628

TABLE 2. Proportion of high weight edges found in the Condensed Matter collaboration network.

p	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.05	8.000	8.000	7.000	2.000	8.000	8.000
0.1	2.667	2.583	2.917	1.167	2.750	3.083
0.15	3.258	2.677	3.032	1.581	2.839	3.516
0.2	2.105	1.698	2.000	1.221	2.070	2.372
0.25	2.141	1.680	1.781	1.234	1.711	2.352
0.3	1.631	1.443	1.552	1.044	1.527	1.847
0.35	1.642	1.335	1.387	1.045	1.348	1.835
0.4	1.536	1.274	1.222	1.121	1.210	1.667
0.45	1.336	1.139	1.113	1.018	1.121	1.450
0.5	1.244	1.094	1.050	1.042	1.055	1.345

TABLE 3. Ratio of high weight links identified in the Condensed Matter collaboration network to those found by random selection.

Observe that for small values of p, the metrics all improve significantly over random selection except preferential attachment, which is by far the worst: it barely performs better than randomness for moderate values of p. The propagation coefficient is superior significant margin over highest common neighbors and Aggregate, and the others are relatively even. Similar effects were also observed in the coauthorship networks for High Energy Physics and Astrophysics, depicted below in Tables 4 and 5, and Tables 6 and 7, respectively.

p	Random	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.05	0.053	0.421	0.421	0.368	0.105	0.421	0.421
0.1	0.126	0.337	0.326	0.368	0.147	0.347	0.389
0.15	0.140	0.455	0.374	0.423	0.221	0.396	0.491
0.2	0.193	0.407	0.328	0.387	0.236	0.400	0.458
0.25	0.225	0.482	0.378	0.401	0.278	0.385	0.529
0.3	0.272	0.444	0.393	0.423	0.285	0.416	0.503
0.35	0.313	0.514	0.418	0.434	0.327	0.422	0.575
0.4	0.334	0.512	0.425	0.408	0.374	0.404	0.556
0.45	0.398	0.532	0.454	0.443	0.405	0.447	0.578
0.5	0.467	0.581	0.510	0.490	0.486	0.492	0.628

TABLE 4. Proportion of high weight edges found in the Higher Energy Physics collaboration network.

p	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.05	8.000	8.000	7.000	2.000	8.000	8.000
0.1	2.667	2.583	2.917	1.167	2.750	3.083
0.15	3.258	2.677	3.032	1.581	2.839	3.516
0.2	2.105	1.698	2.000	1.221	2.070	2.372
0.25	2.141	1.680	1.781	1.234	1.711	2.352
0.3	1.631	1.443	1.552	1.044	1.527	1.847
0.35	1.642	1.335	1.387	1.045	1.348	1.835
0.4	1.536	1.274	1.222	1.121	1.210	1.667
0.45	1.336	1.139	1.113	1.018	1.121	1.450
0.5	1.244	1.094	1.050	1.042	1.055	1.345

TABLE 5. Ratio of high weight links identified in the High Energy Physics collaboration network to those found by random selection.

p	Random	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
.05	0.024	0.404	0.264	0.319	0.030	0.380	0.419
0.1	0.099	0.413	0.268	0.369	0.121	0.432	0.429
0.15	0.140	0.480	0.324	0.399	0.158	0.454	0.492
0.2	0.188	0.470	0.333	0.387	0.193	0.439	0.486
0.25	0.228	0.507	0.355	0.416	0.272	0.454	0.517
0.3	0.285	0.464	0.365	0.401	0.288	0.457	0.483
0.35	0.325	0.497	0.380	0.433	0.336	0.476	0.511
0.4	0.367	0.483	0.395	0.414	0.380	0.463	0.496
0.45	0.413	0.517	0.435	0.444	0.415	0.499	0.530
0.5	0.452	0.499	0.434	0.419	0.431	0.478	0.515

TABLE 6. Proportion of high weight edges found in the Astrophysics collaboration network.

5.4. **Facebook graph.** The dataset is a subset of the New Orleans Facebook network [6]. We are given not only the friendships, but a transcript of wall posts between users during a certain time interval. We constructed a graph from the data of the users and friendships, with all edge weights initialized to 1. We then augmented $w(\overline{xy})$ by 1 if user x posted on user y's wall.

One could argue that this is not necessarily a good indicator of relationship strength.

p	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.05	16.625	10.875	13.125	1.250	15.625	17.250
0.1	4.158	2.695	3.716	1.221	4.347	4.326
0.15	3.421	2.310	2.847	1.130	3.241	3.509
0.2	2.498	1.767	2.057	1.025	2.332	2.579
0.25	2.220	1.554	1.824	1.191	1.988	2.265
0.3	1.629	1.281	1.406	1.011	1.604	1.694
0.35	1.532	1.170	1.333	1.035	1.465	1.574
0.4	1.315	1.075	1.128	1.036	1.259	1.351
0.45	1.251	1.054	1.074	1.005	1.208	1.283
0.5	1.105	0.960	0.928	0.953	1.058	1.140

TABLE 7. Ratio of high weight links identified in the Astrophysics collaboration network to those found by random selection.

Perhaps users that are closest do not interact on Facebook at all, being able to communicate in reality. This is certainly a concern, but in our personal use we have found it to be otherwise. In addition, the paper of Kahanda-Neville [1] shows that wall posts are indeed a strong indicator of relationship strength as judged by "Top Friends" ratings. Nevertheless, it is interesting in and of itself to inquire whether this particular way of weighting can be predicted by intrinsic properties of the graph.

The results for 1000 random nodes of the graph are depicted in Table 8. We tried p in 0.025 increment steps from 0.025 to 0.25. We imposed a cutoff of 2, that is, if $k(x)$ was larger than the number of edges having at least weight two (at least one wall post), we disregarded the excess edges. This was to avoid issues of data corruption arising from the ways in which ties were broken, and explains the slightly non-linear scaling of the random success rate. We also only considered nodes with at least 200 friendships. This is a fairly weak constraint in practice, and we did test without this condition, finding minimal differences in the results.

In Table 9, we show the ratio of the success rate of each metric to the success rate of the random selections: this can be viewed as a "normalized" success rate. The highest common neighbors, Jaccard coefficient, and Adamic/Adar coefficient all perform significantly better than random selection, but preferential attachment does not appear to have any gains over randomness.

p	Random	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.025	0.01539	0.09051	0.10846	0.09271	0.01796	0.10883	0.09930
0.050	0.04349	0.12184	0.14114	0.12765	0.04330	0.14396	0.13158
0.075	0.06547	0.15190	0.16836	0.15678	0.06508	0.17505	0.16244
0.100	0.09377	0.18019	0.19698	0.18526	0.09208	0.20185	0.18953
0.125	0.11662	0.20554	0.22037	0.21095	0.11834	0.22513	0.21537
0.150	0.13310	0.22781	0.23922	0.23171	0.14203	0.24772	0.23659
0.175	0.16040	0.24596	0.25807	0.25208	0.16456	0.26728	0.25612
0.200	0.17578	0.26130	0.27213	0.26718	0.18246	0.28330	0.27017
0.225	0.19178	0.27986	0.28903	0.28383	0.20132	0.29863	0.28673
0.250	0.20365	0.29450	0.30290	0.29855	0.22046	0.31353	0.30103

TABLE 8. Fraction of high weight links identified in the New Orleans Facebook network.

p	Common	Jaccard	Adamic	Preferential	Resource	Aggregate
0.025	5.88095	7.04762	6.02381	1.16667	7.07143	6.45238
0.050	2.80172	3.24569	2.93534	0.99569	3.31034	3.02586
0.075	2.32024	2.57171	2.39489	0.99411	2.67387	2.48134
0.100	1.92161	2.10064	1.97564	0.98199	2.15254	2.02119
0.125	1.76247	1.88967	1.80885	1.01476	1.93043	1.84680
0.150	1.71155	1.79723	1.74082	1.06706	1.86110	1.77754
0.175	1.53341	1.60888	1.57154	1.02594	1.66627	1.59670
0.200	1.48657	1.54817	1.51999	1.03801	1.61173	1.53702
0.225	1.45932	1.50713	1.48001	1.04976	1.55717	1.49511
0.250	1.44607	1.48733	1.46595	1.08250	1.53951	1.47813

TABLE 9. Ratio of high weight links identified in the New Orleans Facebook network to those found by random selection.

We see that the metrics do not improve much over randomness for large p, but become much better as p decreases. In light of the comments made earlier, this is not surprising: we expect the topological considerations to identify very strong relationships, and fall off as the difference between edges becomes less sharp.

We experimented with varying the values of the cutoff in Figure 1. We ran 1000 trials on the Facebook and coauthorship datasets and find that most of the metrics not only perform several times better than random, as seen above, but also continue to improve as this cutoff increases. Preferential attachment, however, did not experience any noticiable changes with an increased cutoff.

6. Discussion

We can see that social networks are heterogeneous. The fact that none of our topological indicators were useful on the Advogato graph speaks to the level of randomness in its edges. On the other hand, we see that topological considerations offer considerable insight into the relationships expressed in the coauthorship networks and Facebook. Our methods were most effective in the Astrophysics network, where we could predict the top edges by a factor of about 16 better than random selection would.

We hypothesize that this disparity stems from the fact that Advogato and Libimseti exist purely as online communities, whereas the Facebook and the collaboration graphs express a network developed in reality. This idea suggests that clustering and common neighbors arise because of geographical or other considerations that affect networks involving physical interaction, and do not arise in online networks facilitating virtual interactions.

Among the different topological metrics, we saw that the simple Most Common Neighbors measure and the Resource Allocation/Propagation were the most effective, with Adamic/Adar and Jaccard's coefficient being slightly worse. In the Facebook network, however, Jaccard's coefficient and Resource Allocation/Propagation outperformed the others. The success of resource allocation in both cases suggests an interesting model of how relationship strength is formed through intermediaries.

It is interesting to reflect on the difference between the Most Common Neighbors ranking and Jaccard's coefficient. The latter is essentially a normalized version of the

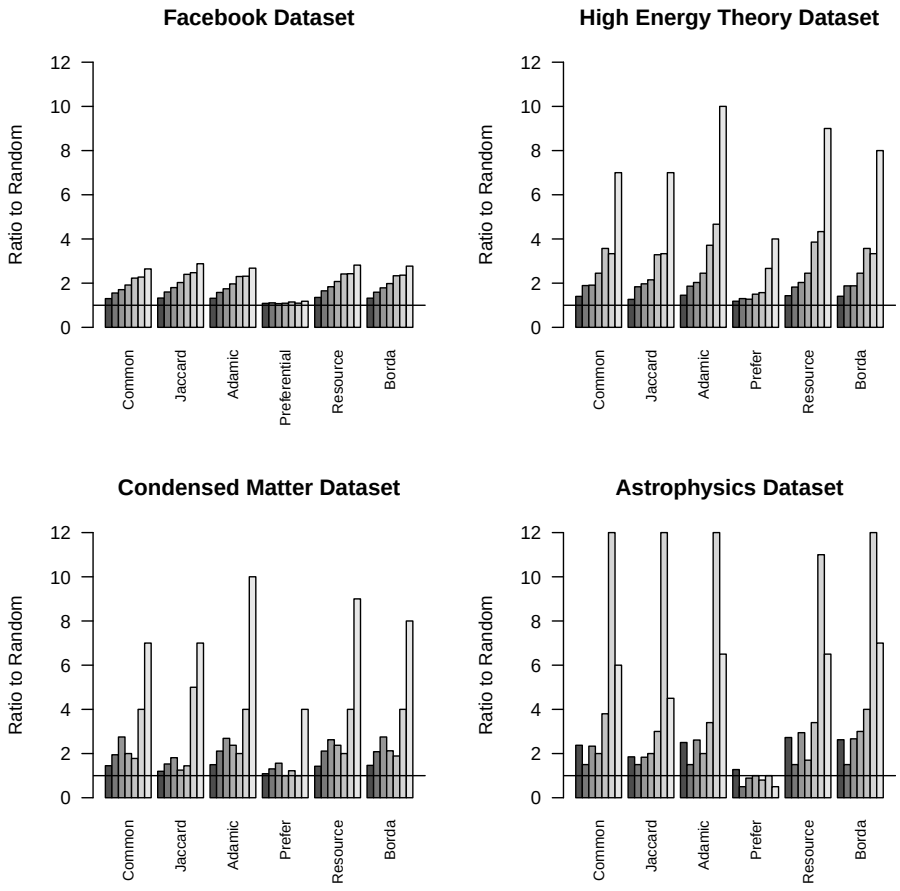

FIGURE 1. Ratio of high weight links to those found in Random for the Facebook and coauthorship networks. Shaded bars represent increasing the cutoff, varying it from 2 to 8.

former. In the collaboration network, it is better not to normalize, but in Facebook it is better to normalize. This result says that people with many collaborators are likely to collaborate many times with any individual one of their collaborators, but people with many facebook friends are less likely to interact much with any individual one of their friends.

The aggregate method hovered around the best of the individual methods, and preferential attachment was uniformly poor. These findings seem to refute the idea that preferential attachment is a good model for the development of relationships in social networks.

We found that the performance deteriorated rapidly as p increased in the Facebook graph, and performance was still sensitive to p in the collaboration networks, but a little less so. That suggests that the structure is mostly among the very well connected nodes, and that the distinction between weaker relationships is essentially random noise.

As we mentioned earlier, iterating the propagation metric did not give better results. Imagining the propagation to represent time spent towards the relationship, we initially hoped that the propagation would model the evolution of relationship strength, and therefore further iterations would improve the accuracy. Unfortunately, this was not the case.

7. Conclusion and Further Directions

We have shown that for certain social networks, exploiting topological structure can yield significant improvements in identifying important links. By ranking links according to certain metrics depending only on the topology of the graph, we were able to predict the top 5% of edges in physics collaboration graphs between 8 and 16 times better than blind guessing, and in Facebook we predicted the top 2.5% of edges about 7 times better than random guessing.

There are many further questions to be pursued. We describe a few below.

- *How can these methods be refined and improved?* Can some rank aggregation scheme be used to combine and improve on all of the individual metrics? We show above that a naive aggregation performs about as well as any of the constituents, but it is conceivable that a more sophisticated method could make further gains. Can our methods be combined with the machine-learning techniques used by other authors, as mentioned earlier, to yield even better results?
- *How does predictability vary among nodes within a given graph?* The variance of performance for these different metrics is high. We found that some nodes are extremely well modeled by our topological measures, and others less so. It would be interesting to study if there are identifiers that indicate whether or not a particular user will be more conducive to these techniques. For instance, if one person's strongest relationships are well-predicted, will his or her friends' relationships also tend to be well-predicted?
- *What type of graph models best represent weighted social networks?* In many cases of this analysis, we found ourselves limited by our datasets, as most of the datasets we found were for unweighted networks. In addition, many datasets were "incomplete." For example, the Facebook dataset represents a small slice of the actual New Orleans network, as the number of friends of a user in this dataset is unrealisitic. Examining and developing generative models to fit the statistical properties of such a network would be useful for future studies.
- *What more can we say about the types of social networks conducive to link estimation?* We have offered a hypothesis concerning networks that have a physical grounding, and networks that exist purely virtually.

We have discussed the potential applications of analyzing the structure of social networks–understanding the interactions between users provides powerful information regarding a particular group of people. As the amount of weighted graph datasets become more abundant, we hope to see these methods generalized to other domains and networks.

References

[1] Indika Kahana and Jennifer Neville. *Using Transactional Information to Predict Link Strength in Online Social Networks.* ICWSM 2009.

[2] David Liben-Nowell, Jon Kleinberg. *The Link Prediction Problem for Social Networks.* Proceedings of the twelfth international conference on Information and knowledge management (2003), 556-559.

[3] Linyuan Lu, Tao Zhou. *Link prediction in complex systems: a survey.* Physica A: Statistical Mechanics and its Applications **390** (2011), 1150-1170.

[4] Paolo Massa, Martino Salvetti, and Danilo Tomasoni. Bowling alone and trust decline in social network sites. In Proc. Int. Conf. Dependable, Autonomic and Secure Computing, pages 658-663, 2009.

[5] M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).

[6] Bimal Viswanath and Alan Mislove and Meeyoung Cha and Krishna P. Gummadi, *On the Evolution of User Interaction in Facebook.* WOSN 2009.

[7] R. Xiang, J. Neville, M. Rogati. *Modeling Relationship Strength in Online Social Networks.* Proceedings of the 19th international conference on World Wide Web (2010), 981-990.