Determining Edge Ranks on a Social Network

Brian Goldman

1 Abstract

We try to predict who a social network user will communicate with the most using only the network's structure. After reviewing prior measures of node-node similarity on a network, we introduce a couple of our own, each based on the intuition that a user's closest friends are more significant to that user and his friends than they are to the whole network. We test these measures on a network with known communication data and find two– the Jaccard index and relative popularity– that significantly predict a user's strongest relationships.

2 Motivation

We know much about the shape of social networks, but often little about the relationships that underlie them. With sufficient computational power, we can traverse the Facebook graph to determine structural data such as the number of friends per user, the graph diameter, and the clustering coefficient. But we compute this data weighing every edge equally, even though the strength of social ties varies wildly. And while Facebook could gauge friendship strength using communication and viewing records, such data is not publicly available. So, we seek to estimate friendship strength using only the structure of the social network and then rank a user's friends according to this estimate.

We consider this a valuable goal because of its broad applicability. Counter-terrorism officials could use a suspect's friendship rankings to distinguish innocuous acquaintances from individuals worthy of greater scrutiny. A brand's marketing team could use a loyal consumer's friendship rankings to determine who that consumer may influence to adopt the brand and who may hardly be exposed to the brand at all.

A better understanding of friendship strength would enhance much of

social network theory. We could design clustering algorithms to group close friends together; we could develop cascading behavior models where consumers prefer the products already used by their closest friends; we could recalculate the distance between two members of a social network using the probability that they communicate with each other. Indeed, researchers have begun to use Facebook data to determine what proportion of attention (wall posts, profile views, etc.) users allocate to their k^{th} best friends (where a user's k^{th} gets the k^{th} most attention).[1]

3 Background

Researchers have proposed several methods to find related members of a network– often in order to recommend products similar to one already enjoyed by a user.[8] Gupte and Eliassi-Rad catalog common approaches.[9] We'll focus on two: the Jaccard index and SimRank.

Given two nodes a and b, let $F(a)$ be the set of a's neighbors, and let $F(b)$ be the set of b's neighbors. Then, the Jaccard index $J = \frac{|F(a) \cap F(b)|}{|F(a) \cup F(b)|}$.[10] Although the Jaccard index is one of the oldest and simplest measures of node similarity, it remains in frequent use. In 2007, Onnela et al. used 18 weeks of national cellphone data to create a social network (where two phone numbers share an edge if each number called the other at least once) and found that two users' Jaccard index strongly correlated with how much time they spent talking to each other.[4] [11]

In 2002, Jeh and Widom proposed SimRank, a fixed-point iterative method to determine node-node similarity. They set a and b's initial similarity score, $R_0(a, b) = 0$ if $a \neq b$ but $R_k(a, a) = 1$ for all k. If $a \neq b$, they define:

$$R_{k+1}(a, b) = \frac{C}{|F(a)||F(b)|} \sum_{i=1}^{|F(a)|} \sum_{j=1}^{|F(b)|} R_k(N_i(a), N_i(b)),$$

where $F(a)$ and $F(b)$ again are a and b's neighbors, respectively. Note that $R_{k+1}(a, b)$ is some constant times the average of $R_k(x, y)$ over all $(x, y) \in F(a) \times F(b)$. The decay rate $C \in (0, 1]$ parameterizes how far similarity spreads from its original source.[1][7]

[1]Jen and Widom set $C = .6$ and $C = .8$, but they suggest that final rankings are not particularly sensitive to this parameter.

Because $R_k(a,b)$ is monotonically non-increasing but bounded above by 1, Jeh and Widom observe that $\lim_{k\to+\infty} R_k(a,b) \in [0,1]$. This limit is $s(a,b)$, or a and b's SimRank similarity score, which satisfies:

$$s(a,b) = \frac{C}{|F(a)||F(b)|} \sum_{i=1}^{|F(a)|} \sum_{j=1}^{|F(b)|} s(N_i(a), N_i(b)).$$

SimRank offers a few distinct disadvantages. At each iteration, we must recalculate the similarity of every two nodes a and b, since $R_k(x,y)$ may be greater than zero for some $(x,y) \in F(a) \times F(b)$ even if a and b do not share an edge. Jeh and Widom suggest "pruning," or only calculating $R_k(a,b)$ for a and b within a small distance of each another.[7] But the process remains slow, and much recent work has focused on finding a quick approximation of $s(a,b)$.[12]

Simultaneous similarity calculation may seem appealing, since it allows us to determine that a and b are good friends both because they share many common friends, and because those shared friendships are strong. But, simultaneous calculation may lead to large error propagation, as misgauging one relationship will lead us misgauge others as well.

We consider one more method, although it has fallen out of use. Let $N(a)$ be the subgraph of a social network restricted to a and its neighbors. The Companion algorithm computes the Hubs and Authorities (HITS) scores for $N(a)$ and then ranks a's neighbors either according to their hub score or their authority score.[5][6]

4 Proposed Algorithms

We will use the Jaccard index and SimRank to try to identify strong relationships on a social network. We also propose a couple algorithms of our own, each based on the intuition that a's closest friends are unusually important within $N(a)$.

4.1 Relative Popularity

We begin by finding the PageRank vector for the whole social network and ranking a's neighbors accordingly.[2] We then find the PageRank vector for $N(a)$ and again rank a's neighbors accordingly. If $b \in F(a)$, let $G(b)$ be b's ranking in the whole network, and $L(b)$ be b's ranking in $N(a)$. We define b's relative popularity $V(b) = \ln G(b) - \ln L(b)$. We hypothesize that a's closest friends will have the greatest relative popularity.

For example, suppose in a given social network, we wanted to rank user 10's friends by relative popularity. User 10 has 5 friends: users 7, 12, 15, 18, and 20. We assign arbitrary PageRank scores to illustrate the process:

User b	PageRank (global)	G(b)	PageRank (local)	L(b)	V(b)	V(b) Rank
7	2.2	1	0.8	2	-0.69	4.5
10	1.4	-	3.0	-	-	-
12	0.9	4	1.1	1	1.39	1
15	0.3	5	.5	3	0.51	2
18	1.2	2	.4	4	-0.69	4.5
20	1.0	3	.2	5	-0.51	3

4.2 Relative Embeddedness

While a pair of acquaintances may happen to know a few of the same people, we suggest that if b and a are close friends, then their mutual friends should form a community. Let $M(a, b)$ be the number of friendships between mutual friends of a and b. If a and b are close friends, then we expect $M(a, b)$ to be unusually large given how many friends b has, $|F(b)|$.

We use the same ranking procedure as before, and again illustrate with an example:

User b	$\|F(b)\|$	$\|F(b)\|$ Rank	M(a,b)	M(a,b) Rank	W(b)	W(b) Rank
7	21	1	7	2	-0.69	3.5
12	12	3	8	1	1.1	1
15	8	4	5	3	0.29	2
18	10	-	0	-	-	-
20	14	2	4	4	-0.69	3.5

where the relative embeddedness, $W(b)$, is the difference of logs of ranks. Note that we exclude user 18 because $M(10, 18) = 0$.

4.3 Comments

Rank data allows us to determine when a user performs better on one measure than another, even if we do not know the underlying probability distributions. We have decided to subtract logs of ranks because doing so accentuates the difference between the lowest (best) ranks. Log differences also exhibit a kind of scale invariance:

$$\ln w - \ln z = \ln(w/z) = \ln(pw/pz) = \ln(pw) - \ln(pz)$$

for any p, so we need not worry over whether to use ranks or percentiles.

The relative popularity method will never suggest a's most (globally) popular friends as candidate best friends. Suppose $G(x^*) = 1$. Then, $L(x^*) \geq G(x^*)$ so $V(x^*) \leq 0$. Likewise, relative embeddedness will never return a's friends who have the greatest number of friends themselves. However, previous research suggests a social media user's best friend is unlikely to be the most popular person that user knows, anyway.[13]

5 Summary of Methods and Results[2]

Using computer science collaboration data provided by Joel Seiferas at the University of Rochester, we created an academic collaboration network, G, where two scholars are connected if they have ever co-authored a paper together with at most one other co-author. Given an author, a, we say b is one of a's top k collaborators if b has co-authored the k^{th} most articles with a, or more (again only counting articles with three or few co-authors). We checked how well the Jaccard index, SimRank, relative popularity, and relative embeddedness replicate these rankings for a sample of 14 authors.

We chose authors with high PageRank scores and many collaborators, because we have the most data about these authors. Because we only have data on author last name, all computer scientists with the same last name form

[2]More thorough results may be found in the appendix.

a single node in G. So, we checked that the selected authors had relatively uncommon names, to increase the chance that they were indeed individual people. As we developed and refined our ranking algorithms, we would regularly test them on a small set of authors. We excluded those authors from the sample.

We are more interested in identifying an author's top collaborators than in ranking his weaker collaborators. So, for each algorithm P, and for each author a, we find the 10 authors P predicts to be most strongly connected to a. Then, we check how many of those authors are among a's top 10 collaborators. The Jaccard index successfully found 48 of our sample authors' 136 top 10 collaborators; relative popularity found 45; relative embeddedness found 24; and SimRank found 12.[3] In fact, the Jaccard index and Relative Popularity made many of the same predictions.

Because of constraints on computational power and time, we only computed SimRank iterations $R_{k+1}(a, b)$ if a and b had collaborated together. We set the decay constant $C = .8$.

We doubted that any graph algorithm could accurately rank an author's weaker ties. Moreover, we cannot even rank weak ties according to number of collaborations, since an author a has worked with most of his collaborators only once. Still, we computed Kendall-τ rank correlation coefficients for each author, comparing the author's true rankings with the rankings predicted by the Jaccard index and relative popularity. The Jaccard index had stronger overall rank correlation for 12 of the 14 authors.

6 Looking Ahead

We have only considered a few ways to identify strong relationships in a social network; other possibilities abound. But even if we cannot devise more effective algorithms, we should try combining the rankings of the best methods that we have already developed. And while our implementation of SimRank performed quite poorly, we would like to test SimRank with less pruning.

We are surprised that the Jaccard index works as effectively as it does. Rather than simply dividing the number of mutual friends by the number

[3]Due to ties, not every author had exactly 10 top 10 collaborators.

of shared friends, we might rank these two quantities and perform a similar rank analysis as we did with relative popularity and relative embeddedness. Unfortunately, prelimary evidence suggests that this is no more effect than just using the Jaccard index.

Our data was a serious handicap throughout this project. Every node in a social network should correspond to a unique person, not a shared name. We should have finer communication data and thus fewer tied rankings. Most of all, our social network should have a predefined notion of friendship. We opted to create an edge between two scholars only if they had coauthored a paper with at most one other collaborator. All our methods relied on clustering– if a and b collaborate often they should have many mutual collaborators. So, if we created an edge between every pair of scholars who had ever collaborated on a paper, papers with many authors would create large cliques, and every relationship within that clique would appear strong. We justify excluding such long papers because we believe many of the collaborators do not actually know each other– but surely some do. Moreover, we wish to be able to find strong relationships on a *given* social network– we cannot simply modify the network at will.

References

[1] Backstrom, L.; Bakshy, E.; Kleinberg, J.; Lento, T.; Rosenn, I. 2011. Center of Attention: How Facebook Users Allocate Attention across Friends.

[2] Lawrence, P.; Brin, S.; Motwani, R.; Winograd, T. 1990. The PageRank Citation Ranking: Bringing Order to the Web.

[3] Flake, G.w., S. Lawrence, C.l. Giles, and F.m. Coetzee. Self-organization and Identification of Web Communities. Computer 35.3 (2002): 66-70.

[4] Onnela, J-P., et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104.18 (2007): 7332-7336.

[5] Dean, Jeffrey, and Monika R. Henzinger. Finding related pages in the World Wide Web. Computer networks 31.11 (1999): 1467-1479.

[6] Kleinberg, Jon M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46.5 (1999): 604-632.

[7] Jeh, Glen, and Jennifer Widom. SimRank: a measure of structural-context similarity. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[8] Small, Henry. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24.4 (1973): 265-269.

[9] Gupte, Mangesh, and Tina Eliassi-Rad. Measuring tie strength in implicit social networks. Proceedings of the 3rd Annual ACM Web Science Conference. ACM, 2012.

[10] Jaccard, Paul. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr. Corbaz, 1901.

[11] Easley, David, and Jon Kleinberg. Networks, crowds, and markets. Cambridge Univ Press 6.1 (2010): 6-1.

[12] Yu, Weiren, et al. A space and time efficient algorithm for SimRank computation. World Wide Web 15.3 (2012): 327-353.

[13] DeScioli, Peter, et al. Best Friends Alliances, Friend Ranking, and the MySpace Social Network. Perspectives on Psychological Science 6.1 (2011): 6-8.

A Complete Results

We may use the hypergeometric cumulative distribution function to determine a significance level for our top 10 collaborator results. For example, the relative popularity method found 7 of node 1185's top 10 collaborators. If the relative popularity method randomly returned 10 of user 1185's 118 collaborators, the probability of finding at least 7 top 10 collaborators is $1 - CDF(6, 118, 10, 10) \approx 2.5\text{E-}7$.[4]

[4]In the tables below, x is the author index. $|F(x)|$ is x's total number of collaborators. J, P, E, and S show the number of top 10 collaborators correctly identified by Jaccard, relative popularity, relative embeddedness, and SimRank. And α is the calculated significance level.

Author x	$\|F(x)\|$	J	α_J	P	α_P	E	α_E	S	α_S
Rao 311	127	2	.15	1	.53	0	1	0	1
Alon 1185	118	4	4.4E-3	7	2.5E-7	4	4.4E-3	2	.20
Naor 3552	123	3	3.5E-2	5	2.4E-4	1	.58	0	1
Papadimitriou 1395	93	6	2.1E-4	6	1.0E-4	3	9.4E-2	1	.74
Tarjan 1817	84	3	9.4E-2	3	9.4E-2	1	.54	2	.34
Das 1857	92	4	1.1E-2	4	1.1E-2	1	.70	0	1
Berman 2963	78	3	.14	3	.14	2	.43	2	.43
Melhorn 3072	78	4	1.3E-2	4	1.3E-2	1	.73	2	.32

Of the 14 authors included in our experimental sample, we initially tested only 8. Because this was such a small group, and because Jaccard and relative popularity top 10 collaborator results were so similar on this group, we decided to test the other 6 authors as well. The first 8 are above, the last 6 are below.[5]

Author x	$\|F(x)\|$	J	α_J	P	α_P	E	α_E	S	α_S
Fischer 711	105	4	2.5E-3	0	1	0	1	0	1
Agrawel 1330	92	1	.66	2	.25	2	.25	1	.66
Shamir 1342	96	5	4.3E-4	3	4.7E-2	2	.24	1	.65
Klein 1650	87	2	.37	1	.76	1	.72	0	1
Wagner 1720	134	5	1.6E-4	5	1.6E-4	5	1.6E-4	1	.55
Simon 2989	99	2	.22	1	.63	1	.63	0	1

Finally, we calculated the Kendall-τ rank correlation coefficients.

[5]Due to ties, Rao, Melhorn, Agrawel, Shamir, and Simon had 9 top 10 collaborators. Papadimitriou, Berman, and Klein had 11 top 10 collaborators. Fischer had 8 top 10 collaborators. Jaccard predicted 11 top 10 collaborators for Papadimitriou. Relative embeddedness predicted 6 top 10 collaborators for Tarjan and 9 top 10 collaborators for Klein.

Author x	Jaccard τ	Relative Popularity τ
Rao 311	.306	.053
Alon 1185	.230	.218
Naor 3552	.249	.259
Papadimitriou 1395	.400	.166
Tarjan 1817	.335	.196
Das 1857	.290	.223
Berman 2963	.150	.122
Melhorn 3072	.203	.074
Fischer 711	.194	.027
Agrawel 1330	.227	.039
Shamir 1342	.321	.334
Klein 1650	.164	-.048
Wagner 1720	.210	.157
Simon 2989	.289	.191