Mind Economy: Social Capital in Social Networks

Alexy Khrabrov George Cybenko
Thayer School of Engineering, Dartmouth College
{alexy,gvc}@dartmouth.edu

October 28, 2010

Abstract

Social networks such as Twitter and Facebook increase their mind-share daily, and many online activities are determined by the interactions there. In order to understand the dynamics of social networks ("what's happening"), we want to identify the key players – those participants who are in some way important, influential, and possess certain social capital. Although these terms are used in sociology and computer analysis of static networks, we need to come up with the new and more rigorous definitions in the face of the new users getting on Twitter daily, making the network inherently dynamic. Our solution is a set of dynamic metrics of importance, called D-Rank and StarRank, which allow ranking over time and in comparison to one's audience (network neighborhood). Using these metrics, we uncover fascinating worlds inside Twitter, such as the Justin Bieber ecosystem and Brazilian sport journalists with their fans. Building upon the insights from those communities, we define our version of Karmic Social Capital as an iterative update rule, rewarding those who facilitate balanced and stable communication. Running a complete world emulation with our rules, we end up with the social capital distribution which places the hard-talking "middle class" nodes at the top, leading to a new ranking and understanding of the Twitter dynamics.

1 Introduction

Given a large social network changing over time such as Twitter, how do we find the key people who are important, influential in some sense, and have certain social capital – and how do we formally define and measure these terms? Computer and social scientists, economists, operation researchers, and educators have all proposed quantitative and qualitative approaches for describing social capital. As a consequence, the term "social capital" is so widely [over]used since the 1980s that some researchers shy away from it altogether [4]. Furthermore, the three qualities listed often overlap. We define them usefully and distinctly for

our networks, applying these new definitions to find interesting phenomena such as fan-based economies of ratings and trends, and multi-modal collaboration of hackers, advancing open-source projects online via social coding.

The modern economy is knowledge-based, and knowledge generation closely corresponds to value and wealth creation. The processes where knowledge is created, refined, and made actionable, are increasingly shifted into social networks on the Internet, embedded in social media sites such as Twitter and Facebook, or underlying cooperation on Wikipedia, the social coding portal GitHub, etc. Certain members and groups are key to the value-creating processes in these networks, where people are united by the information they work with. Those individuals contributing the most and best knowledge, or processing it in the best way, are recognized as the main contributors, and get to set the agenda for the whole group. In fact, a lot of communications are questions for the senior members, seeking explanation or coordination. It is important to quantify which members gain importance in these mind economies to know how they work.

The quintessential mind economy is the programmers' community. Their industry, spawning startups in the Bay Area, Pacific Northwest and elsewhere, trades in ideas becoming code becoming startups becoming large companies like Amazon and Google. (In turn, the geeks' ethos epitomized by these companies reflects back on the society where they thrive.) The resources traded by hackers are most commonly URLs of code repositories on GitHub, the open-source social coding portal. GitHub is organized around Git, a distributed source code management system (SCM) authored by Linus Torvalds and originally used to develop the Linux kernel. Now it is a de facto platform to collaborate on open-source projects. While GitHub is used to store and modify the code, people working on it often converse on the Internet Relay Chat (IRC) and Twitter. Some of the most active geek communities on Twitter center on advanced programming languages such as *#scala*, *#clojure*, and *#haskell* – these are the "hash tags" used to mark tweets so that they can be found as a group. Coincidentally, these are also the names of the corresponding IRC channels.

Another community with high traffic is what we call Justin Bieber ecosystem, discovered and described in [7]. Justin Bieber is a boy pop-singer phenomenon, originating on the YouTube and spawning an intense following there and on Twitter. Many of his fans are teenage girls with a variation of "bieber" in their Twitter nick and are united in their adoration of @justinbieber and pushing him up into the top 10 trending topics on Twitter. In this process, the members learn to increase their own ratings by swapping shoutouts (mentions) and trading shoutouts for follows. An economy of rating-increasing behavior develops. Those who get followed by Justin Bieber increase their standing among the "beliebers" immensely, as do those who organize other fans around better schemes "how to meet Justin." High-intensity group behaviors are key in social dynamics and change, and can be studied for the first time on a coherent social organism of a Twitter community with its drivers and influencers, and the methods they use to direct it.

2 Approach

In this thesis, we approach the problem of importance via dynamic analysis of the communication graphs. We build a communication graph of mentions, where one user talks to another via public tweets. E.g., if @alice tweets: "@bob did you see this: http://bit.ly/xyz", she *replies* to @bob, who is thus her *replier*, while she is his *mentioner*. While many Twitter analyses and ranking sites focus on the number of followers [5], we prefer communication as a form of active behavior with its many social implications, manifesting itself similarly in social media networks, email networks, and real life.

Traditional sociology measures betweenness and centrality on static graphs [1]. Their networks are small and their tools are often Excel spreadsheets and add-ons. By contrast, our data subset consists of a 100 million tweets by 5 million users, with both numbers growing over the period of 35 days. (In turn, our subset is selected from the "gardenhose," a Streaming API for a "statistically representative" fraction of Twitter, spewing now 5 million tweets daily, from which we collected billions of tweets.) We have to take temporal nature of these data into account as a key feature of the model.

2.1 Importance

First, we propose a measure of importance we call Dynamic PageRank, or D-Rank [7]. For every day in the study, we treat the communication graph as a directed multigraph (as Alice can easily tweet Bob thrice day) and compute the PageRank of all the nodes present that day. We translate such PageRanks into relative ranks, from 0 to 1, showing where the node stands in the overall sorted list of ranks.

We then define StarRank as a ratio of a node's D-Rank to the average D-Rank of its communication partners. There are also variations where we count a node's repliers and his mentioners either together or separately. Both D-Rank and StarRank can be compared across days, even though the number of nodes on Twitter increases every day.

2.2 Social Capital

We propose a mathematically well-defined measure of Social Capital, which we call Karmic Social Capital, which accrues for those dialogue participants who better maintain their question-answering balance in conversations – they do reply to those who address them, and get replies from those to whom they talked before. This capital can also favor stronger ties, where you keep talking to your current interlocutors, or it can reward exploration of new partners. The model is parameterized with weights (rewards) for different kinds of behaviors, and previous capital decays with time.

Most definitions of Social Capital are in fact just other kinds of importance measures. E.g. Getoor [9] simply uses that to denote the number of your co-authors on conference program committees. A good review of the definitions

of social capital used in computer science is provided in [10]. We model our *Karmic Social Capital* on the social capital of Tuscany villagers who remember how many times they lent everyday items like salt to each other, versus how many times they got even by borrowing some anchovies, or more salt, in return. A good working description of this kind of local social capital is provided in e.g. [3]. Every community member in those social-capital-rich areas has a clear mental balance of the favors given and received, which figures in every subsequent social and economic transaction.

Our first computational model is "karmic" repayment of communications – instead of a carefully maintained balance of favors, we have a communication network where you repay questions (mentions) by replying (mentioning in return). There's a temporal notion of capital accumulation, along with its decay in the face of inaction.

3 Related Work

When looking for influence in social networks, several classes of problems turned out to be closely related to our definition of influence in a community. We addressed them in the additional papers on which this thesis will also build.

In [8], we considered the question of network structure which enables the networks to withstand random faults or malicious attacks, taking out some nodes one by one. It is one of the first papers which studied the malicious attacks on a network, and compared behavior of differently structured networks, such as scale-free or random, under different destructuring scenarios. It is an application of dynamic graph analysis, examining how the influential nodes can help keep the network together.

In addition to Twitter, we studied a sensor-based social network, resulting from the MIT Reality experiments [2]. A fundamental question in dynamic systems is the agents' identity. We address it in [6], where we were able to identify a majority of the MIT Reality participants from just about 10 hops in their cell phone traces. Social importance of the subjects is related to their patterns of action (motion) and resulting interactions.

4 Karmic Social Capital

Definitions. *Repliers* of a node are the addressees of its tweets. *Mentioners* of a node are those who tweet to it. If a tweet from *@Alice* mentions *@Bob*, then *@Bob* is a replier of *@Alice* and *@Alice* is a mentioner of *@Bob*. Replier of a node is someone that that node repliers *to*. In other words, from a node's perspective, repliers are out-degree, mentioners are in-degree.

symbol	definition
S_v^t	Social Capital of node v at time t. Superscript t generally denotes "by time t." Specifically *during* time step t is denoted as $@t$.
$G^t(V, E)$	graph G with nodes V and edges E
$w_{uv}^{@t}$	total weight of directed edges $u \to v$, i.e. the number of tweets from u to v during time step $@t$
W_{uv}	total number of undirected edges between u and v: $W_{uv} = w_{uv} + w_{vu}$
B_{uv}	Balance of back and forth tweets from u to v: $B_{uv} = w_{uv} - w_{vu}$
M_u	$\{v \mid w_{vu} > 0\}$, i.e. the mentioners of u
$R_u^{@t}$	$\{v \mid w_{uv}^{@t} > 0\}$, i.e. repliers of u specifically during the timestep $@t$
$O_{uv}^{@t}$	outgoing activity of a node rewarded by social capital at timestep t
$A_{uv}^{@t}$	incoming activity in this cycle rewarded just for mentions (all)
$B_{uv}^{@t}$	incoming mentions in this cycle repaying previous replies (balance)
α, β, γ	model parameters

$$
\begin{aligned}
O_u^{@t} &= \frac{1}{\sum_{V^{t-1}} O^{@t-1}} \sum_{v \in M_u^{t-1} \cap R_u^{@t-1} | B_{uv}^{t-1} < 0} |B_{uv}^{t-1}| w_{uv}^{@t-1} W_{uv}^{t-1} S_v^{t-1} \\
B_u^{@t} &= \frac{1}{\sum_{V^{t-1}} B^{@t}} \sum_{v \in M_u^{@t-1} | B_{uv}^{t-1} > 0} B_{uv}^{t-1} w_{vu}^{@t-1} W_{uv}^{t-1} S_v^{t-1} \\
A_u^{@t} &= \frac{1}{\sum_{V^{t-1}} B^{@t}} \sum_{v \in M_u^{@t-1}} w_{vu}^{@t-1} W_{uv}^{t-1} S_v^{t-1} \\
I_u^{@t} &= \gamma B_u^{t-1} + (1-\gamma) A_u^{t-1} \\
S_u^t &= \alpha S_u^{t-1} + (1-\alpha)(\beta O_u^{t-1} + (1-\beta)(\gamma B_u^{t-1} + (1-\gamma) A_u^{t-1}))
\end{aligned}
$$

Some notes on the definitions. O_u^t is the node u's output gaining social capital, thus we want to reward those who redress an imbalance of input and answer those who addressed you more than you had answered them. The summation is defined exactly over those with whom you have a deficit in replying: $v \in M_u^{t-1} \cap R_u^{@t-1} | B_{uv}^{t-1} < 0$. It means the target node mentioned you at some point prior, you replied to it in this cycle, and before that, you owed it a reply since it tweeted more to you than you did to it. For each such deficit node you replied to, finally, we multiply the number of replies in that cycle, $w_{uv}^{@t-1}$, by the balance you owed, $|B_{uv}^{t-1}|$, the value of the relationship, W_{uv}^{t-1}, and the importance of the replier S_v^{t-1}. We normalize the O's so that they all sum to 1, and reward each node proportionally to the value of the redress in the reply imbalance it actively contributed in this cycle.

Similarly, I_u^t is the node u's input worth of social capital. We generally consider all input as good – we can't distinguish bad publicity, or consider it all good anyways – but we distinguish mentions redressing the mentioners' own deficit with us as worthing more than just any mention. Those repaying

mentions we reward with a multiplier for the balance owed additionally to the usual cycle contribution, relationship value, and the mentioner's social capital.

Note that in the output, we don't have a general activity term for all replies, even those not redressing an imbalance, as we do in the second term of the input. Thus we don't reward random replying, and you won't get social capital by just addressing everybody in volume.

It's easy to see that such a definition of social capital allows for an iterative economy by launching the update rule defining S_u^t in terms of S_u^{t-1}as shown in the last formula above.

5 Fundamental Questions

The core questions, which may allow for more quantitative treatment with our metrics:

1. Why do people twitter? What is the utility and can it be captured by a form of social capital? Can a single definition suffice for all members of a network? Do "beliebers" differ from hackers and how, w.r.t. their forms of social capital and behaviors increasing it?

2. How can you increase your social capital or importance in the fastest, but most robust ("honest," irreversible) way? How can we distinguish fake importance from the real thing?

6 Data Mining

Given our metrics of importance and social capital, we explore interesting individuals from our large Twitter data set and interpret what these metrics mean in real life.

The first results of our full-world Social Capital emulation show a new kind of nodes at the top, in addition to the usual suspects such as Justin Bieber (still in the top 10) – we call it "the middle class of social networks." These are the people with a certain amount of followers – though not as many as the stars – e.g. from a 100 to a 1,000, weaving a web of intense dialogues with other such nodes. They provide the core of the ongoing communication, pushing trends and stars up, bringing uninteresting subjects down by not talking about them, integrating other nodes and beginners into the system. We believe our metrics are more interesting than the traditional PageRank in that they reflect the dynamic nature of the conversational networks and highlight the groups making lasting ongoing contributions.

A typical example is beauty industry. Two owners of Singapore beauty salons come on top in terms of continuous conversation threads kept the longest. They discuss wedding photos, haircuts, design, etc. They have a constant and necessary market, thousands of followers each, and discuss news of their industry. This way they keep on top of their market, satisfy their followers with specific goals, and stay on top in the ratings.

Hence one of the differences in the utility of tweeting. Celebrities, those most followed in a typical definition, often get on top by projection from the real world, YouTube, etc. They don't have to tweet for their supper, so to speak. Their top fans and other in-Twitter user phenomena have to work for it. While spammers achieve a short-term notoriety by mass following, hoping for automatic follow-backs, we do not register those numbers, since we look for real conversations and mentions. Top celebrities like @*donniewahlberg* and @*justinbieber* understand this and cultivate their fan base, encouraging top fans once in a while. The timing feature of our metrics smooth over the spikes typical in social networks and reveal true value.

7 Conclusions and Future Work

We considered metrics of importance and the fundamental questions they should answer. We introduced Karmic Social Capital, a money-like measure which decays with time and rewards socially beheficial behavior. Using this metric, we discovered the "middle class" of Twitter – those whose conversations carry most of the steady discourse. The next steps will explore various forms of this social capital, such as subtracting from the givers and adding to the takers proportionally to their current capital, going further into transitive transfers, and more general theory of such exchanges.

8 Bibliography

References

[1] P. Bonacich. Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.

[2] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Pers Ubiquit Comput*, 10(4):255–268, May 2006.

[3] Dario Gaggio. *In Gold We Trust: Social Capital and Economic Change in the Italian Jewelry Towns.* Princeton University Press, 2007.

[4] M.O. Jackson. *Social and economic networks.* Princeton University Press, 2008.

[5] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[6] Alexy Khrabrov and George Cybenko. A language of life: Characterizing people using cell phone tracks. In *Proceedings IEEE CSE'09, 12th IEEE*

International Conference on Computational Science and Engineering, Vancouver, BC, Canada, pages 495–501, 2009. August 29-31.

[7] Alexy Khrabrov and George Cybenko. Discovering influence in communication networks using dynamic graph analysis. *Proceedings IEEE CSE'10, 13th IEEE International Conference on Computational Science and Engineering, Minneapolis, MN (to appear)*, 2010.

[8] Alexy Khrabrov, David M Pennock, C. Lee Giles, and Lyle H Ungar. Static and dynamic analysis of the internet's susceptibility to faults and attacks. *INFOCOM 2003*, Sep 2003.

[9] Louis Licamele, Mustafa Bilgic, Lise Getoor, and Nick Roussopoulos. Capital and benefit in social networks. *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, Aug 2005.

[10] BKD Motidyang. *A Bayesian belief network computational model of social capital in virtual communities. Ph.D. Thesis in Computer Science.* University of Saskatchewan, 2007.