

Discovering Influence in Communication Networks using Dynamic Graph Analysis

Alexy Khrabrov and George Cybenko
{alex,y,gc}@dartmouth.edu
Thayer School of Engineering
Dartmouth College
8000 Cummings Hall
Hanover, NH 03755

Abstract—The rise of Internet-based social networks has shifted many decision-impacting discussions online. Increasingly, people weigh new ideas, choose products, pick technologies, find entertainment and socialize virtually by engaging in online discourse. The participants depend on who people find online, who they get to know and trust, and who they consider as authorities on subjects of interest. This paper presents techniques to track who has influence in such a network and how they got there. Many definitions of influence are possible; here we focus specifically on the social interaction and its dynamics, using Twitter as the reference network and data source.

We build a replier graph from each user A 's messages mentioning another user B (which may be either “for” or “about” B), and study how this graph evolves. (In a tweet from A mentioning B , A is the replier mentioning B .) For every day in the study, we compute a pagerank-type score and a *drank*, a dynamic function of the pagerank, for all users, together with a series of features such as the number of mentions a user gives or receives. The daily-versioned features enable exploratory data analysis of the conversational dynamics by looking at the relative decline or growth in specific features for every user every day, separately or relative to others. For instance, we find the longest periods of growth in the number of times a user A is mentioned by other users on a day d , $m = |M(A, d)|$, over a contiguous period of days, and also compute its acceleration over that period, dm/dt . Those accelerating the most, or sustaining the longest growth, or both, are worth closer modeling.

Our metrics are applicable to any evolving directed graphs and allow us to find people of growing influence in social networks based purely on the structure and dynamics of their conversations. These are the first dynamic metrics for social networks which take into the account both global and local influence (pagerank and repliers), and can be applied to other communication networks as well. Most interestingly, using them, we uncover a high-intensity ecosystem with its own “mind economy,” adapting to maximize the participants’ rankings and promote their shared message.

I. INTRODUCTION

The web becomes more and more interactive and real-time every day. The original static Mosaic pages were replaced by database-driven merchants and blogs. An early pioneer of web-based conversation, LiveJournal, appeared in 1999 and still has a majority mindshare in countries like Russia and such communities as programmers and journalists, due to the ease of dialogue. Comments became *de rigeur* for blogs, and now new social media such as Twitter and Facebook are essentially comment and status update-driven. Comments

and status updates are inherently two-way communications, and these social networks may supplant SMS and email, becoming full-fledged communication networks. Traditional communication networks, such as email, still dominate, and in many cases require analysis, such as has been done on the Enron data set. When presented with such a dataset, an important question is, who’s important here and who’s not? Whose influence should we discern behind the dynamics of the network to understand its processes?

While forming a social circle on Twitter, people often follow those who are authorities on their professional or personal interests. Since most of the experts now publish their Twitter *@nickname* on the Web, it is easy to follow them. In many cases, people ask questions on Twitter as they used to do on blogs, as on LiveJournal, asking for advice on products to buy, technologies to use, travel destinations, restaurants, etc. Advice is often sought in conversations, and decision are arrived at via consulting the subject or social authorities. An influential microblogger can make or break a company’s reputation, speed up technology uptake, and accelerate a trend or do the opposite. Thus it’s important for those in the public arena to understand who wields influence in the social networks, how they get there, and how this influence evolves over time.

In this paper, we study conversational dynamics influence in communication network graphs. We use Twitter as a publicly available communication network. We look at the communication graph formed by messages, or tweets, from one person mentioning another, those replies (or mentions) mapping to an edge in the graph. We then look at several graph metrics dynamically, computing them for every day in the study, and then look at the users whose metrics dominate, or accelerate faster, or grow longer than others’. It turns out that such users form interesting classes, and one of our metrics, the *starrank*, can show when a user becomes a public persona.

We believe that this work can be used in a variety of ways: individuals can track changes and dynamics of their influence as well as their peers’ influence; competitive intelligence analysts can analyze the dynamics of consumer opinions about their products and services with a deeper understanding about who has key influence about attitudes and opinions; intelligence analysis of an organization’s evolution and changing roles within an organization.

TABLE I
Drank COMPUTATION

[a:0.1 b:0.02 c:0.3 d:0.2]	original user:pageranks
[b:0.02 a:0.1 d:0.2 c:0.3]	sorted by pageranks ascending
[b:0 a:1 d:2 c:3]	pageranks replaced by position, 0-based – this is the <i>dirank</i>
[b:0 a:1/4 d:2/4 c:3/4]	normalized by the list length – this is the <i>drrank</i>

II. PREVIOUS WORK

Java, Finin, et al., [1] look at the power users of Twitter and find that they are indeed superheroes, their energy activated by more replies. They emphasize the difference between the declared network of followers and the actual network of social interactions. Backstrom and Kleinberg [2] studied community growth in two social networks, LiveJournal and DBLP (computer conferences). They looked at a few temporal snapshots, learned some rules of community growth in between, and found some of the graph structures which make such processes as joining a community more likely. Khrabrov and Cybenko [3] apply sequence modeling to the cell phone tracks of a group of MIT students, using n-gram models and suffix trees. Savell [4] and Chung [5] have developed techniques for identifying and analyzing business process dynamics within social networks using a variety of methods including Process Query Systems [6].

III. DATASET AND METHODOLOGY

We started monitoring Twitter using its new Streaming API from June 2009 onward. Our subscription level, the so-called “gardenhose,” provides a fraction of all tweets, which is quite significant: about 2-3 million tweets a day initially, now reaching 4-5 million tweets daily. The working set for the studies in this paper consists of a 100 million tweets from October 16 to November 17, 2009. From those, we compute our metrics on the leading 90 million tweets, spanning three complete weeks (22 days to disregard time zone effects).

The replier graph is built from this working set in an incremental way. For each new day, we add that day’s mentions as directed edges, and count them for the source (in number of repliers), and the target (in the number of mentions). We compute the pagerank [7] of each user at the end of the day, using the Jung2 toolkit [8] with $\alpha = 0.15$, over 100 iterations. Given the pagerank for each user, for each day, we define the *drank* of every user that day as follows:

- Sort the $(user, pagerank)$ pairs by increasing pagerank;
- Replace each user’s pagerank by its position (starting from 0);
- Optionally normalize into 0-1 range by dividing by the list’s length.

Table I contains a step-by-step example of the *drank* computation.

The number of users with a given pagerank increases daily, as more and more users get on Twitter and engage in conversation. The daily number of ranked users is shown in Figure 1.

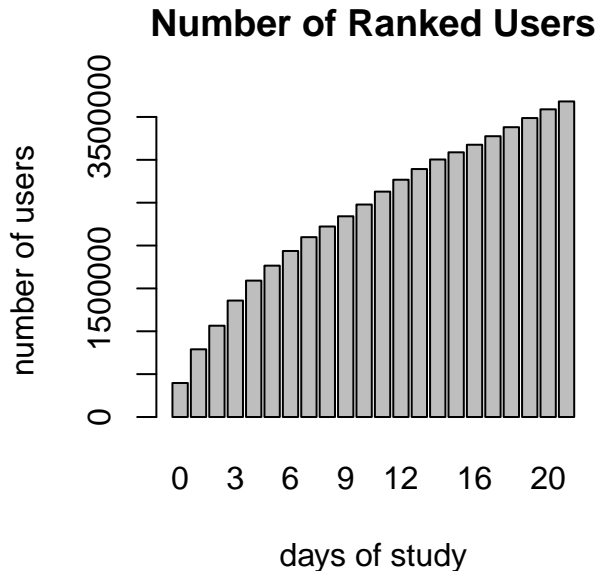


Fig. 1. Twitter’s user base continues to grow significantly

Normalization compensates for the ever growing number of users active on Twitter every day. However, the effects where such growth is interesting are better observed on the integral *dranks*. We refer to the integral, non-normalized *drank* as *dirank*, and the rational, normalized one as *drrank*. Simple metrics such as the number of mentions a user gave or received each day are integers, which, in the dynamic contexts, become integer lists. We look at the number of repliers (different targets mentioned by the same source), mentions (different replies with the same target), and total number of tweets by a user, daily, which translate into integer lists.

Those histograms contain only those users whose complete corpus is nondecreasing or increasing; a single daily drop will filter them out. To look at all users, we partition their lists into longest contiguous nondecreasing or increasing subsequences, and compute the maximal length and/or acceleration of such a subsequence. The acceleration is a ratio of the last element over the first; since many first elements are 1, we also look at a “tougher” version, where the acceleration is with respect to the second element; the minimal length of the subsequence in each case is 3, showing at least two days of growth (or no decrease).

Many features related to influence increase or decline along with time, and are represented as a real-valued time series. Examples are the number of mentions a user receives daily, his *drrank*, and *drstarrank*. Totals such as number of mentions given or received can be counted separately for each day or cumulatively from the beginning of the study. Below we describe several primitives which we combine for our analyses.

- *Contiguous Longest Increasing (generally Monotonic) Subsequences* – This operation (*clis*) takes a sequence,

TABLE II
CONT. LONGEST INCREASING SUBSEQUENCES

[1 2 3 0 4 5 8]	original sequence
[[1 2 3][0][4 5 8]]	subseqs partitioned by $<$, $maxlen$ 3
$\begin{bmatrix} 3 & 0 & 8 \\ 1 & & 4 \end{bmatrix}$	subseq accelerations, $maxxel$ 3/1, $maxxel-tough$ 8/5

TABLE III
GROW OR FALL

[1 2 3 5 0 7 6]	original sequence
5 pairs up, 2 pairs down	:grow both simple and qualified
rate of change is 6/1	passes :twice-higher filter

an ordering function – one of $<$, \leq , \geq , $>$ and partitions it into subsequences such that each subsequence is ordered according to the predicate. Acceleration for each subsequence is the ratio of the last element to the first; the *:tough* variant divides by the second instead, skipping the frequent 1 for temporal count sequences. We can drop the leading 0s also when computing accelerations, or filter for them by multiplying the ratios out against a threshold. We denote the maximum length of a *clis* subsequence as *maxlen*, and the maximum acceleration as *maxxel* (or *maxxel-tough*). Table II gives an example of the *clis* transform.

- *Grow or Fall* – This operation (*growfall*) takes a sequence and counts how many successors are greater or less than their predecessors. Then a decision is made whether the sequence is mostly “grow” or “fall” by either the simple (1 : 1) or qualified majority (2 : 1). Optional *twice-higher* filter keeps only those sequences where the change between the first and last element is twice or more. The result can be either categorical – *:grow*, *:fall*, or *:neutral*; or quantitative – the rate of change, with 0 for neutrals, and the sequence returned ordered by the rate of change. Table III shows an example of the *growfall* transform.

The *starrank* considers a user’s importance with respect to his neighborhood. Figure 2 shows a user with his mentioners on a particular day, along with his and their *dranks*. The *starrank* is an average of the neighbors’ ranks, weighted by the number of communications with each neighbor for that day. Depending on the kind of the d-rank we use, we’ll get a corresponding *starrank*. For instance, a user’s *distarrank* for a day will be an average of the *diranks* of her neighbors, i.e. of their positions in that day’s sorted order of page ranks.

IV. RESULTS

For individual ranks sequences we can look at the longest growth or decline periods. We partition each sequence into contiguous increasing or decreasing subsequences and take the longest one. For the number of mentions, we look at the total amount of users for whom it is nondecreasing, and look at their longest runs over days. Figure 3 shows how many users with all non-decreasing mention runs have full sequences of at least a given number of days. Each bucket is the total number of users who achieve that many days of stability or

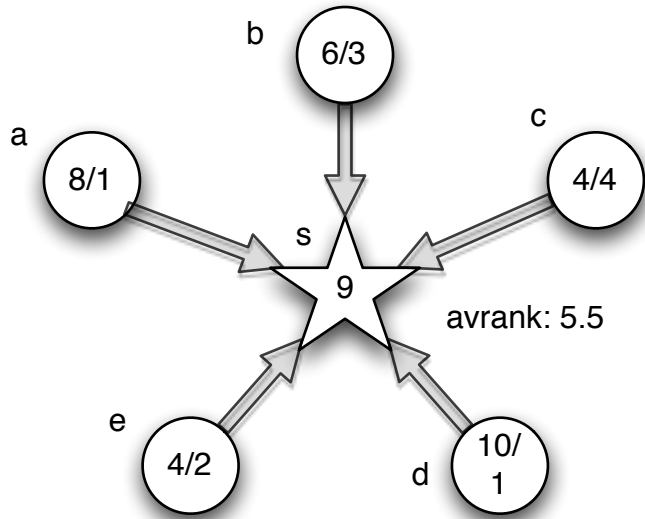


Fig. 2. Example of *starrank* computation. The center user with *drank* of 9 is mentioned by 5 other users with the given *dranks*, exchanging one or more tweets that day, e.g. $r/n = 6/3$ means 3 mentions by a user of *drank* 6. Then the *distarrank* is the average of *rs* weighted by the *ns*, here $(8 * 1 + 6 * 3 + 4 * 4 + 10 * 1 + 4 * 2) / (1 + 3 + 4 + 1 + 2) = 5.5$

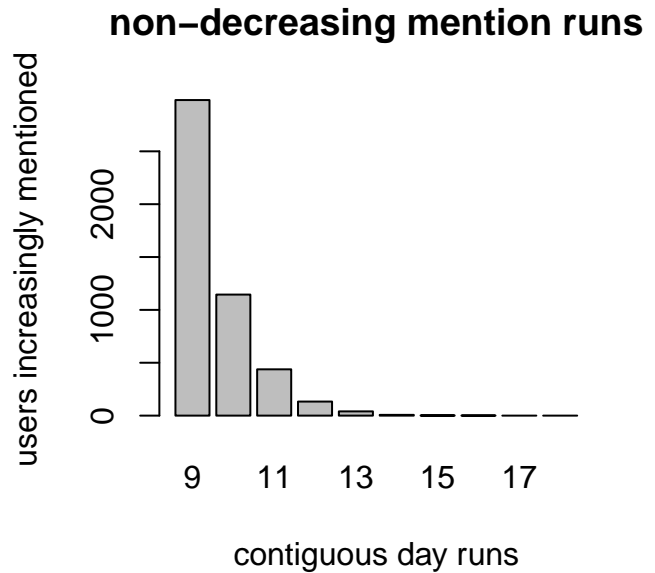


Fig. 3. The number of users whose daily mentions are all nondecreasing, per day

growth. Figure 4 shows the same for non-increasing *dirank*. The one person who persists to the end is Justin Bieber with the top nondecreasing rank of 0 (see below on the Bieber “Ecosystem”).

The users with the longest period of growth in their *drrank* include the following:

- Brazilians. All of our growing replier metrics bring up a

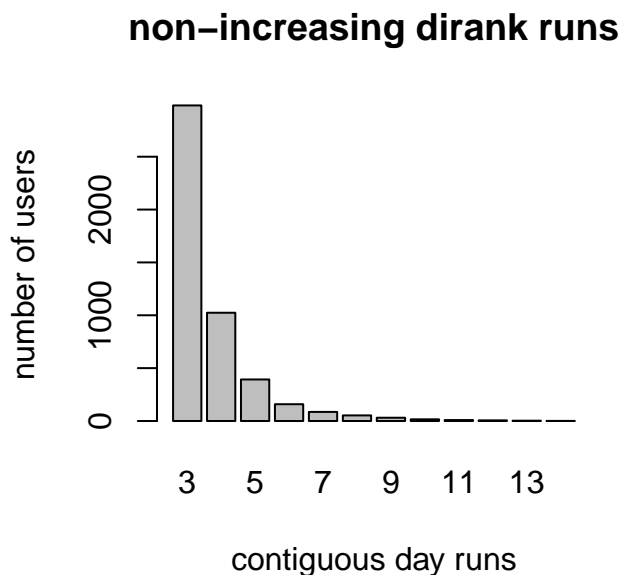


Fig. 4. The number of users whose own daily *diranks* are all nondecreasing, per day

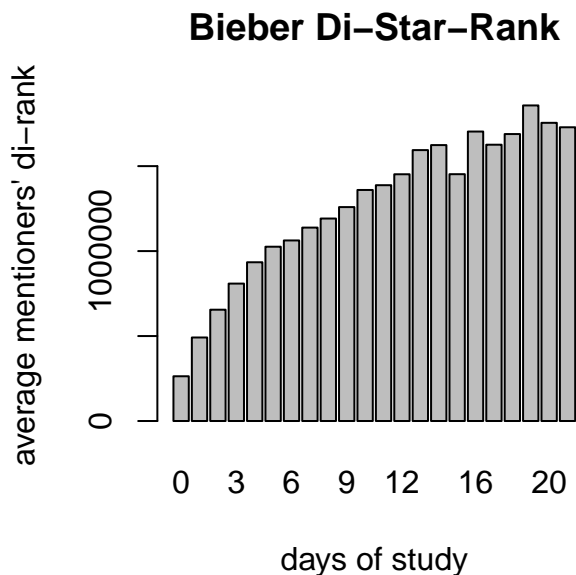


Fig. 5. The average *dirank* of Justin Bieber’s fans, decreasing daily, shows his star power spreading to the masses

lot of Brazilian journalists, stars, and power users. Apparently, Brazilian twitterers are more communicative than others, using Twitter more like a conversation medium than a diary. The top ones include @leobarcellos, a web designer, @natyperdomo, “Publicitária,” and more.

- @alfiehitchcock – a London photographer posting his pictures via TwitPic
- @Theresamcardle – an “Optimist; Toilet seat marketer at Kohler; UW-Madison MBA student”

Those with the longest contiguous increase in real *starrank* of mentioners – i.e. losing influence – include such accounts as @gm_web, an automatic repost of a weather station, and @AlexanderFog – a self-referential DJ referring to himself in every tweet.

The second-order characteristic of growth is acceleration. Here, for each contiguous increasing or decreasing subsequence, the acceleration is given by the ratio of the final state to the original value. For a quantity such as the real rank, where growth means decrease, we invert it, so that the bigger acceleration, the better the growth is. Once again, the pack of *starrank* acceleration-sorted users is dominated by many of the same Brazilians, showing that the longest active growth also often is the fastest one. Some of the new faces here are

- @jc_schuster – a country music fan
- @drosa_shannon – “twin,married, love DOOL,BL,American Idol and Tweeting”
- @carlaciccarelli – “Publicitária e Produtora de Comerciais,Vídeos e Eventos”
- @Dollbabyv – a girl writing about her dating, referencing a Blogspot blog

The *starrank* clearly demonstrates how a star constantly spreads its influence through to ever increasing realms of new users, with lower *dirank*, thus its average audience rank is increasing. Figure 5 shows the daily *distarrank* of Justin Bieber, the most influential replier. It’s not normalized to show the effect of the influx of the new users.

The normalized *starrank*, i.e. *drstarrank*, when decreasing for the longest runs over days, shows the cumulative growth in the importance of a community around the user with respect to all others. When sorted by such a longest run, we get interesting classes of influential users, such as

- @donniewahlberg – leader of the New Kids on the Block band and fans, with far-reaching and active following
- @bowwow614 – the rapper Bow Wow
- @faydra_deon – “Minister; Computer Applications Trainer; Website Designer; WordPress/Joomla! Customizer; Grammar Queen; Online Bookstore Owner; AKA,” posting her own “questions of the day.”

Now we look at the acceleration of the *starrank*, and find the users with the fastest growing communities by importance. Some of the top users with the accelerating mentioners are:

- @JoycePascowitch – “jornalista, glamourosa....e estressada!”, a Brazilian journalist
- @Biofa – “Jornalista cruzeirense que ama futebol, mulher e rock’n roll (meu Deus, como isso é bom!!!!)”, a Brazilian sports journalist
- @Dejdia – apparently a photographer in LA, using a Flickr page as the URL, with her own ecosystem of fans
- @Minni_w – “ddub Soldier (Army of NKOTB)”, a fan of @DonnieWahlberg, one of the top-page-ranked users, the

band leader for the New Kids On The Block (NKOTB). She connects with the Bieber network (see below).

The snapshots of the pagerank relationship to number of mentions reveals certain stable patterns. We cluster the pagerank vs. number of mentions graph into groups of 1000 points, and take the median of x s and y s to plot, which we call a blocked projection. Figure 6 shows how the resulting graphs form a “harp” pattern which stays stable through days 10 and 20. We filter on x to stay within 25 median mentions on a day, the remaining few clusters group the outliers with exceptional dynamics (or statics).

A similar pattern persists for the relationship between the pagerank of a user and the sheer number of tweets by that user. A block projection in Figure 7 shows that twittering more gets you only so far, and then your ranking starts to fall again. We show it for day 20, day 10 is similar.

V. THE PAGERANK GENERATION

The most surprising discovery in the Twitter network we made is the Bieber ecosystem. Its growth principle can be summarized as: *You can't be Justin Bieber, but you can be Amanda D.* The person with pagerank 0, i.e. the most influential person in Twitter mentions in our dataset, each day, every day of the study, is Justin Bieber. Not many people over age 15 know who he is, despite his Christmas 2009 performance at a Washington, D.C. concert for President Obama. An informal polling of seminar participants and friends over 18 confirms an almost absolute lack of Bieber awareness. At the same time, Justin Bieber, a YouTube boy singer phenomenon, commands a significant mindshare among the pop culture fans from about 11 to 15 years old, from Brazil to Slovenia, but mostly teen girls in Canada and the US.

The most interesting aspect of the Bieber Ecosphere is that it's the first Twitter generation which grows up with social networks in hand, and adapts to its influence seeking ways almost naturally. They evolve to efficiently rise in the page ranks. The second and third daily position by pagerank are most often occupied by several high-caliber fans of Justin Bieber. One of them, Amanda D., self-billed as a singer and songwriter in the South, is too young to drive, and her parents wouldn't take her to a Justin's concert in the nearest big city, so one of her online goals is to get him to perform in her hometown – as described on a free blog website with a link offered to Justin and others to share. What's most fascinating of the self-named “beliebers” is their active and incessant manipulation of the pagerank by constantly mentioning each other and growing the followers' network.

Mentioning another fan is called a “shoutout,” which are offered for trade on a quid pro quo basis – “SHOUTOUTS FOR SHOUTOUTS!” Multiple accounts are created for special purposes, e.g. @Wewanttomeetjustin is maintained by the two top *beliebers*, for the purpose of aggregating other fans who want to meet Justin, discussing venues, options, plans, etc. User @BieberFame has created a separate @JBieberCash account, noting its creation date and that Justin followed the same day, and openly referring to the original account as the

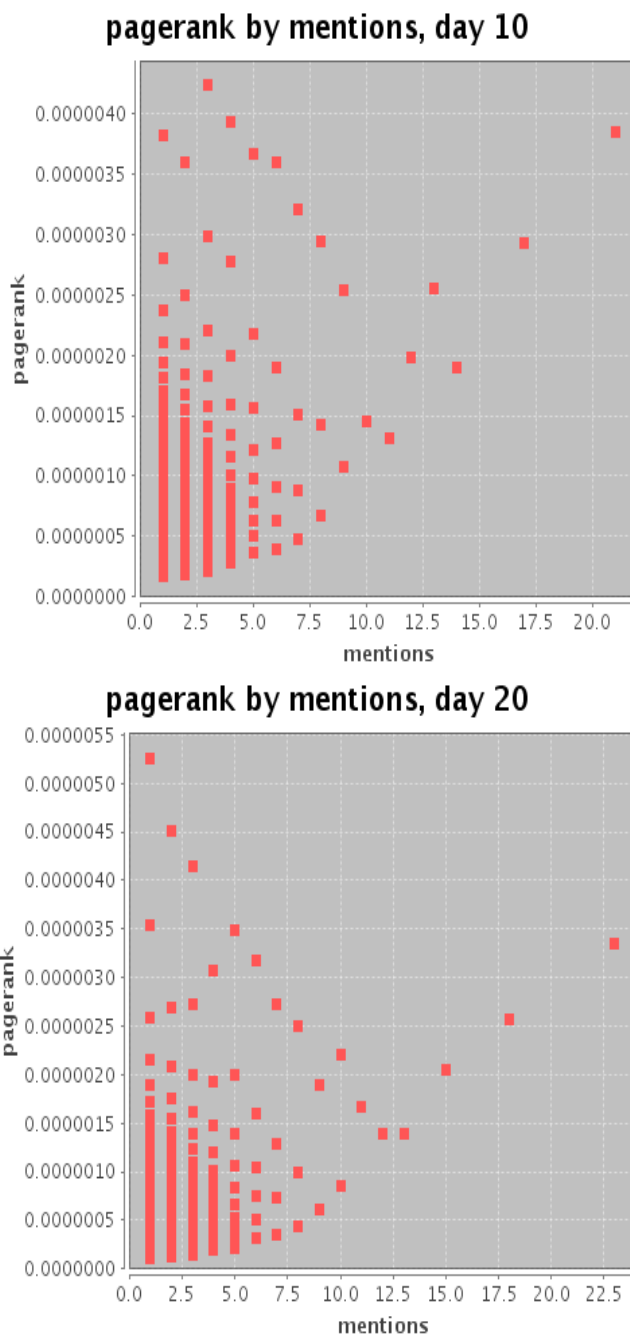


Fig. 6. Pagerank improves with the number of mentions only so much, then ratcheting mentions is counterproductive. The harp pattern persists throughout the days

pagerank by numtwits, day 20

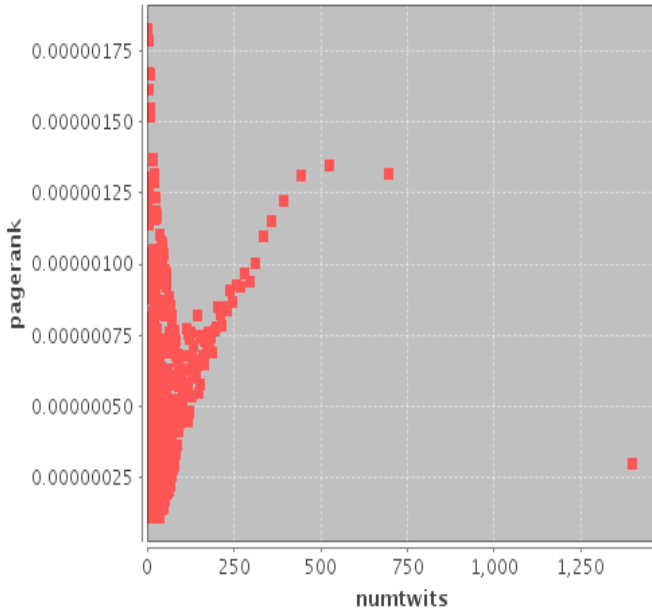


Fig. 7. Pagerank improves with the number of twits only so far as well. X is the cumulative number of twits, Y is the resulting pagerank

creator. Very common tweets state that the number of followers is almost near e.g. 10,300, with only about 50 remaining, so all new followers will get a follow-back and a shoutout. Many tweets are simply multiple shoutcasts, i.e. lists of names.

The top *beliebers* got where they are by active and clever Twitter presence, increasing their influence in a variety of ways. They trade shouts and create multiple Twitter accounts for focused subgroups, tending them regularly, and team with other top *beliebers* to do it, positioning themselves at the head of the pack. They also claim to maintain special relationship with Justin, direct-messaging with him, and offer to relay other fans' question via this exclusive channel. As far as we can see, these advantages are real – Justin mentioned both two top *beliebers*, thus propelling their pageranks to the top. The *starrank* pattern of the top fans resembles that of the stars themselves, showing a steady increase of mentions and increase of the average rank of the mentioners, propagating “into the masses.” Amanda is so popular and effective that she got her own fan account, “Amanda D’s Army,” in turn run by one of the top-level Bieber fans using the same techniques which proved to work so well with Bieber’s fans.

@*Kekeinaction* shows up high in the list of the longest contiguous growth in *distarrank* of mentioners, on almost all days of the study (21). She maintains heavy cooperation with a lot of other Bieber fans, and also was a star of the movie “*Akeelah and the Bees*.” This shows how the younger segment of the audience, focused on teenage culture, is densely connected. Another example is a top fan from Slovenia, @*jbieber_fever24*, discovered at 7th position of the longest *distarrank* decreasing runs. Her bio reads, “Hey I’m Tea I’m

from Slovenia and I’m 14.I LOVE Justin Bieber!I also like Tay Swift,Miley Cyrus and Sel Gomez!Justin followed me 25/10/2009!” She uses all the tricks in the book – trading shoutouts, ramping up the number of followers, maintaining multiple accounts such as @*JBieber_Babes*, and inviting other fans to participate in all of those activities together.

VI. CONCLUSIONS AND FUTURE WORK

In our study we applied dynamic graph analysis to the communication graph of Twitter repliers. We developed metrics for individual users and also their communities which measure the growth of influence over days, by finding the longest runs of growth in mentions, ranks, *starranks*, and their accelerations. We found that the *starrank* is a good indication of influence, and uncovered a whole segment of the Twitter population naturally engaged in rank promotion by increasing the density of the dynamic communication graph, creating multiple accounts, raising the number of followers and repliers, etc. We also found actively growing national communities, namely in Brazil, and identified types of online personae – journalists and musicians – which quantifiably generate the most engagement from the audience in the social network. All those metrics are based purely on the graph and temporal structure of the communications. In our future work, we will use the actual text context of the messages exchanged and n-gram modeling to further quantify the dynamics of influence in communication over social networks.

VII. IMPLEMENTATION NOTES

The massive amount of streaming social network data makes the technology choices crucial. In this research we rely on the JVM platform and one of its functional programming languages, *Clojure*, a dynamic Lisp [9], taking advantage of the scalability and parallelism it offers. We receive Twitter data via its Streaming API as JSON and store it in *MongoDB* [10], a modern *NoSQL* document database. The data mining is performed in *Clojure*, interfacing with *MongoDB* via *congomongo* [11]. Any intermediate results can be stored right back as nested maps of maps or vectors in *congomongo*, allowing for transparent serialization, persistence, and indexing via *MongoDB*. A graph with decorated edges is represented as a *Clojure* map, which, along with a vector, are first-class constructs – e.g. a graph of mentioners from the Figure 2 looks like

```
{:s {:a [8 1] :b [6 3] :c [4 4] :d [10 1] :e [4 2]}}
```

Such maps are subject to *Clojure* destructuring and functional map/reduce/filter transforms, expressive and concise. The code for our project is open source and available on github [12]. Visualization and model fitting is smoothly handled by *Incanter*, an R-like statistical environment in *Clojure* [13].

VIII. ACKNOWLEDGMENTS

The authors thanks Thayer School at/and Dartmouth College for providing real communication networks; Alexy thanks the staff of the Rosey’ Cafe in Hanover, NH, for the science

enablement with coffee, Andrew Boekhoff for congomongo, David Liebke for Incanter, and the 10gen team for the excellent MongoDB and fantastic support (think 5 minutes github turnaround).

REFERENCES

- [1] A. Java, X. Song, T. Finin, and B. L. Tseng, "Why we twitter: An analysis of a microblogging community," in *WebKDD/SNA-KDD*, ser. Lecture Notes in Computer Science, H. Zhang, M. Spiliopoulou, B. Mobasher, C. L. Giles, A. McCallum, O. Nasraoui, J. Srivastava, and J. Yen, Eds., vol. 5439. Springer, 2007, pp. 118–138.
- [2] L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *KDD*, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. ACM, 2006, pp. 44–54.
- [3] A. Khrabrov and G. Cybenko, "A language of life: Characterizing people using cell phone tracks," in *CSE (4)*. IEEE Computer Society, 2009, pp. 495–501.
- [4] R. Savell and G. Cybenko, "Mining for social processes in intelligence data streams," *Social Computing, Behavioral Modeling, and Prediction*, Edited by Huan Liu, John J. Salerno and Michael J. Young, 2008.
- [5] W. Chung, R. Savell, J.-P. Schtt, and G. Cybenko, "Identifying and tracking dynamic processes in social networks," *Proceedings of SPIE, Vol. 6201, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V*, Edward M. Carapezza, Editors, 620105, 10 May 2006.
- [6] G. Cybenko and V. H. Berk, "Process query systems," *Computer*, vol. 40, no. 1, pp. 62–70, 2007.
- [7] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [8] J. O'Madadhain, D. Fisher, and T. Nelson, "Jung – a free and open-source java software library for manipulating, analyzing, and visualizing network data," <http://jung.sourceforge.net/>.
- [9] R. Hickey, "Clojure – the lisp that makes the jvm dynamic," <http://github.com/richhickey/clojure/>.
- [10] I. 10gen, "Mongodb – a fast caching nosql document database," <http://mongodb.org/>.
- [11] A. Boekhoff, "Congomongo – a clojure wrapper for the mongodb java api," <http://github.com/somnium/congomongo/>.
- [12] A. Khrabrov, "Conversation graph mining toolkit," <http://github.com/alexymongol/>.
- [13] D. E. Liebke, "Incanter – a clojure-based, r-like platform for statistical computing and graphics," <http://incanter.org/>.