# Strathprints Institutional Repository

Laflin, Peter and Mantzaris, Alexander Vassilios and Grindrod, Peter and Ainley, Fiona and Otley, Amanda and Higham, Desmond J. (2012) *Dynamic targeting in an online social medium.* In: SocInfo 2012, 2012-12-05 - 2012-12-07, Lausanne.

# Dynamic Targeting in an Online Social Medium

**Abstract.** Online human interactions take place within a dynamic hierarchy, where social influence is determined by qualities such as status, eloquence, trustworthiness, authority and persuasiveness. In this work, we consider topic-based Twitter interaction networks, and address the task of identifying influential players. Our motivation is the strong desire of many commerical entities to increase their social media presence by engaging positively with pivotal bloggers and tweeters. After discussing some of the issues involved in extracting useful interaction data from a Twitter feed, we define the concept of an *active node subnetwork sequence*. This provides a time-dependent, topic-based, summary of relevant Twitter activity. For these types of transient interactions, it has been argued that the flow of information, and hence the influence of a node, is highly dependent on the timing of the links. Some nodes with relatively small bandwidth may turn out to be key players because of their prescience and their ability to instigate follow-on network activity. To simulate a commercial application, we build an active node subnetwork sequence based on key words in the area of travel and holidays. We then compare a range of network centrality measures, including a recently proposed version that accounts for the arrow of time, with respect to their ability to rank important nodes in this dynamic setting. The centrality rankings use only connectivity information (who Tweeted whom, when), but if we post-process the results by examining account details, we find that the time-respecting, dynamic, approach, which looks at the follow-on flow of information, is less likely to be 'misled' by accounts that appear to generate large numbers of automatic Tweets with the aim of pushing out web links. We then benchmark these algorithmically derived rankings against independent feedback from five social media experts who judge Twitter accounts as part of their professional duties. We find that the dynamic centrality measures add value to the expert view, and indeed can be hard to distinguish from an expert in terms of who they place in the top ten. We also highlight areas where the algorithmic approach can be refined and improved.

## 1 Motivation

Centrality measures have proved to be extremely useful for identifying important players in an interaction network [27]. Although the fundamental ideas in this area were developed to analyse a single, static network, there is a growing need to develop tools for the *dynamic* case, where links appear and disappear in a time-dependent manner. Key application areas include voice calls [9, 14], email activity [3, 14], online social interaction [29], geographical proximity of mobile device users [17], voting and trading patterns [1, 25] and neural activity [4, 12].

This work focuses on the use of centrality measures to discover influential players in a dynamic Twitter interaction network, with respect to a given topic, with the aim of finding suitable targets from a marketing perspective. In this social interaction setting, the idea of key players, who influence the actions of others, is intuitively reasonable. Emperical evidence is given in [11] for *discussion catalysts* in an on-line community who are "responsible for the majority of messages that initiate long threads." Further, Huffaker [16] identifies *on-line leaders* who "trigger feedback, spark conversations within the community, or even shape the way that other members of a group 'talk' about a topic.". Experiments in [24] on email and voice mail data found evidence of individuals "punching above their weight" in terms of having an ability to disseminate or collect information that cannot be predicted from static or aggregate summaries of their activity. These people were termed *dynamic communicators*, and an explanatory model, based an inherent hiererchy among the nodes, was suggested. Such concepts make it clear that the dynamic nature of the links plays a key role—the *timing* and *follow on effect* of an interaction must be quantified if key players are to be identified. A recent business-oriented survey [6, Section 4] lists network dynamics as a key technical challenge, and the authors in [28] argue that "the temporal aspects of centrality are underepresented."

Several recent articles have addressed the issue of discovering important or influential players in networks derived from Twitter data. The work in [2] focused on how a shortened URL is passed through the network. Using the premise that a person who passes on such a URL has been influenced by the sender, it studies the structure of cascades. Related work in [23] looked at large scale information spread on the Twitter follower graph in order to measure global activity. The authors in [8] studied a large scale Twitter follower graph and compared three meaures that quantify types of influence: number of followers (out degree), number of retweets and number of mentions, finding little overlap between the top Tweeters in each category. Similarly, [22] also ranked users by the number of followers and compared with ranking by PageRank, finding the two measures to be similar. By contrast, they found that the retweet measure produces a very different ranking. We note that none of the influence measures considered in [8, 22] fully respect the time-ordering of Twitter interactions. For example, reversing the arrow of time does not change the count of followers, retweets or mentions. In this sense, they overlook a crucial aspect of the interaction data. Our work differs from that described above by (a) focussing on subject-specific Tweets of interest in a typical business application, (b) building the interactions between Tweeters on this topic and recording them in a form that we call the active node subnetwork sequence, and (c) comparing a range of centrality measures in this dynamic setting, including one that respects the arrow of time, against independent hand curated rankings from social media experts exposed to the same data.

## 2 Building the Active Node Subnetwork Sequence

The Twitter business home page at `https://business.twitter.com/basics/what-is-twitter/` explains that

> "Anyone can read, write and share messages of up to 140 characters on Twitter. These messages, or Tweets, are available to anyone interested

> in reading them, whether logged in or not. Your followers receive every one of your messages in their timeline—a feed of all the accounts they have subscribed to or followed on Twitter. This unique combination of open, public, and unfiltered Tweets delivered in a simple, standardized 140-character unit, allows Twitter users to share and discover what's happening on any device in real time. ”

The number of active Twitter users currently exceeds 140 Million, with over 340 Million Tweets generated per day. Of direct relevance to our work, the business home page adds that

> “Businesses can also use Twitter to listen and gather market intelligence and insights. It is likely that people are already having conversations about your business, your competitors or your industry on Twitter. ”

Twitter is a means to send out information over a well-defined network. This brings to life a scenario that social scientists have for many years been using as a theoretical tool to develop concepts and measures. In particular, given only a network interaction structure, perhaps describing social acquaintanceship, it has proved extremely useful to imagine that information flows along the links and thereby to identify important actors [10, 27]. In this setting, most centrality measures are defined through, or can be motivated from, the idea of studying random walks along the edges [26], or deterministically counting geodesics, paths, trails or walks [7]. These ideas have been extremely well accepted and widely used, despite the obvious simplifications that the methodology involves. For example, even if we accept that social acquaintanceship is a reasonable proxy for the links along which information flows, there are issues concerning

**link types:** if A and B are acquainted professionally and A passes on some work-related news to B, then it is reasonable to expect that B is more likely to pass this news on to professional colleagues than other friends. So we could argue that some A→B→C paths have a greater chance of being traversed than others.

**link dynamics:** if A and B meet only on a Sunday evening, and B and C meet only on a Monday morning, then we could argue that even though the undirected path A ↔ B ↔ C exists in the network, the route A→B→C is a more likely conduit for news than C→B→A. This is because B meets C soon after an A→B exhange, and hence is more likely to (a) remember and (b) regard as topical, any information received from A. This gives another sense in which paths are not created equal.

By exploiting features of the Twitter data, we can, to some extent, sidestep the shortcomings above while retaining the elegance and simplicity of the network-based view:

**link types:** each link represents a physical exchange of information that is known to have taken place (rather than a proxy such as social acquaintanceship), and moreover, by filtering based on Tweet content, we can, in principle, record only links that are relevant to a specific topic of interest,

**link dynamics:** the Twitter data gives us access to the time at which each piece of information was disseminated.

Twitter's follower graph, where nodes represent users and a directed link connects a user to a follower, has been studied, for example, in [8, 22, 23]. In our work, we wish to focus on users who are engaging with a particular topic, so a natural first step is to look at those who send Tweets containing a predefined set of phrases. In principle, the followers of all such users are exposed to the information in those Tweets. However, in practice we do not know if or when a follower reads a Tweet or acts upon it outside the Twitter platform. In this work, we focus on clearly *active* nodes, that is, users who send out at least one Tweet on the required topic. We then focus on directed user-to-follower connections that involve these active nodes. As well as ruling out those Tweets that land on 'stony ground' this pruning exercise generally has the effect of reducing the size of the network considerably; an issue that is of importance if we wish to consider global Tweets about popular topics over long time scales.

To be precise, we use the Twitter feed to construct an *active node subnetwork sequence* as follows.

**Definition 1** *The* active node subnetwork sequence*:*

- *Start the clock at time* $t_{\text{start}}$
- *Listen to all Tweets that contain the required phrase(s)*
- *Each time a new Tweet is recorded, make sure the sender and all the sender's followers are nodes in the network (i.e. add them if necessary), and add a time-stamped directed link from the sender node to all follower nodes.*
- *Stop the clock at time* $t_{\text{end}}$
- *Post-process the network by removing all nodes that have zero aggregate out degree, i.e., remove those people who did not send out any relevant Tweets.*
- *Slice the data into* $M$ *windows of size* $\Delta t = (t_{\text{end}} - t_{\text{start}})/M$*. We will let* $t_k = t_{\text{start}} + (k-1)\Delta t$*. Then, for* $k = 1, 2, \ldots, M$*, the* $k$*th window covers the time period* $[t_k, t_{k+1}]$ *and is represented by an integer-valued matrix* $A^{[k]}$*. Here* $(A^{[k]})_{ij}$ *records the number of links from node* $i$ *to node* $j$ *that appeared in this time period.*
- *Binarize each* $(A^{[k]})_{ij}$*, that is, set all positive integers to the value* 1*. (See the remark below for a discussion of this step.)*

Implicit in this definition is the simplifying assumption that a Tweet has an influence over a fixed period of time, $\Delta t$. It may be argued that a Tweet, once sent, exists for ever and should create a permanent link that perpetuates across all subsequent time windows. However, we believe that a more compelling argument is that Tweets are time-sensitive and fairly rapidly disappear down a typical follower's timeline. The choice of $\Delta t$ then quantifies the typical "read and respond" time.

We emphasize in particular that reducing $\Delta t$ does not necessarily give a more accurate representation of reality—although we know the precise time that the Tweet was sent, we do not know if or when each follower digests the content.

On the other hand taking $\Delta t$ too large (e.g. one giant window) causes us to lose information about the time-ordering of the Tweets.

We constructed an active node subnetwork sequence by listening to Tweets containing the phrases `city break`, `cheap holiday`, `travel insurance`, `cheap flight` and two phrases relating to specific travel brands. This simulates a typical client-driven investigation on behalf of a travel company wishing to improve its social media presence. The collection took place from 17 June 2012 at 14:41 to 18 June 2012 at 12:41. We took $\Delta t$ equal to 66 minutes, producing 20 time windows. The total number of Tweeters and followers associated with this data set is $442,948$. Restricting attention to active nodes, with nonzero out degree, reduced the network size to $N = 590$.

We observed that some accounts can Tweet a lot in a short space of time. One account Tweeted 104 times in timeframe 10 and a further 23 times in timeframe 11. This account released a total of 127 Tweets in 68 minutes. This motivates our decision to binarize the data within each window—in this way we have not taken account of how many times an account Tweets, but rather we represent the fact they did Tweet in that timeframe. This is done to try to stop the overall result being influenced by accounts using a high volume of automated Tweets. This choice is a balance between allowing a "noisy" account broadcasting automated Tweets to score higher than we would like in our calculations against our ability to pick out influential people by observing a natural increase in the rate of conversation because something interesting or relevant is happening.

To give a feel for the data, Figure 1 visualizes two portions of the the network at the end of the first time window.

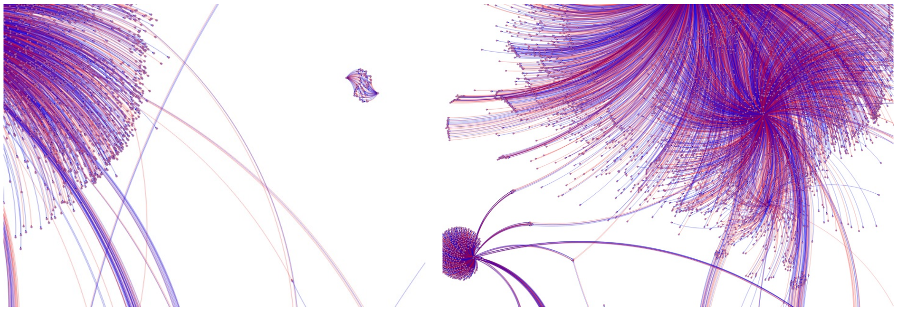

**Fig. 1.** Two details from the active subnode network sequence at the end of the first time window; in particular, showing the existence of an isolated community.

We will return to this data set in section 4 when we compare centrality meaures.

## 3 Centrality Measures

In the case of a single time point, with binary adjacency matrix $A \in \mathbb{R}^{N\times N}$, the resolvent matrix $(I-\alpha A)^{-1}$ was proposed by Katz [18] as a means to summarize pairwise "influence" under "attenuation through intermediaries." Here the fixed parameter $\alpha$ governs the strength of the attenuation, and for $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ denotes the spectral radius of $A$, we have

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \ldots$$

Using the fact that $(A^p)_{ij}$ records the number of distinct walks[1] of length $p$ from node $i$ to node $j$ [10], we see that the $(i, j)$ element of $(I-\alpha A)^{-1}$ counts the total number of walks of all possible length, with walks of length $p$ downweighted by $\alpha^p$. The idea of attaching less importance to longer walks is intuitively reasonable, and Katz [18] also points out that $\alpha$ may be intepreted probabilistically, as the chance that a message successfully traverses an edge. It follows that the row sums and column sums of the resolvent quantify the ability of nodes to broadcast and receive information, respectively. Rather than inverting $I - \alpha A$, it is more efficient and numerically accurate to solve a linear system. Hence in our tests we will compute vectors Kb and Kr in $\mathbb{R}^N$ satisfying

$$(I - \alpha A)\mathrm{Kb} = \mathbf{1}, \qquad (I - \alpha A^T)\mathrm{Kr} = \mathbf{1}, \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector with all entries equal to one. In this case the $i$th components of Kb and Kr measure the ability of node $i$ to broadcast and receive messages, respectively, across the static network represented by the binary adjacency matrix $A$, in the sense of Katz. The nodes may then be ranked according to these scores.

In the limit $\alpha \to 0$, longer walks make a negligible contribution in (1), and, ignoring uniform shifts and scalings that do not alter the rankings, the measures collapse to out degree and in degree, respectively, that is,

$$(\mathrm{deg}_{\mathrm{out}})_i = \sum_{j=1}^{N} a_{ij}, \qquad (\mathrm{deg}_{\mathrm{out}})_j = \sum_{i=1}^{N} a_{ij}. \tag{2}$$

We note that these two quantities are also widely used as centrality measures in their own right [10, 27].

In recent years, several authors have pointed out that concepts such as geodesics, paths and walks can be extended to the case of a time-ordered sequence of networks [5, 15, 19–21, 25]. We focus here on the dynamic walk notion from [14] which produces generalizations of the Katz centrality measures (1) that are feasible for large-scale network computations. Using the notation introduced in section 2, the following definition was made in [14].

[1] A walk of length $w$ from node $i$ to node $j$ is characterized by a sequence of $w$ edges $i \to i_1$, $i_1 \to i_2$,..., $i_{w-1} \to j$. There is no requirement for the edges, or the nodes that they connect, to be distinct.

**Definition 2** *A* dynamic walk of length $w$ *from node* $i_1$ *to node* $i_{w+1}$ *consists of a sequence of edges* $i_1 \to i_2, i_2 \to i_3, \dots, i_w \to i_{w+1}$ *and a non-decreasing sequence of times* $t_{r_1} \leq t_{r_2} \leq \ldots \leq t_{r_w}$ *such that* $A^{[r_m]}_{i_m, i_{m+1}} \neq 0$.

Dynamic walks are easily counted by forming appropriate matrix powers. For example, with the $(i,j)$ component relating to walks from node $i$ to node $j$,

- $A^{[1]}A^{[2]}$ counts all dynamic walks of length two that use one edge at time $t_1$ followed by one edge at time $t_2$,
- $A^{[3]}A^{[4]}A^{[6]}$ counts all dynamic walks of length three that use one edge at each time $t_3$, $t_4$ and $t_6$, in that order.
- $A^{[5]}A^{[5]}A^{[9]}A^{[10]}$ counts all dynamic walks of length four that use two edges at time $t_5$, and then an edge at time $t_9$ and finally an edge at time $t_{10}$.

Following the Katz idea of downweighting walks of length $w$ by $\alpha^w$, this leads to the expression

$$\left(I - \alpha A^{[1]}\right)^{-1} \left(I - \alpha A^{[2]}\right)^{-1} \cdots \left(I - \alpha A^{[M]}\right)^{-1}$$

as a summary of the number of dynamic walks that exist between each pair of nodes. In this case, $\alpha$ should be chosen below the reciprocal of $\max_{1 \leq k \leq M} \rho(A^{[k]})$.

Expressing these computations in terms of sparse linear systems, rather than matrix inversions, and normalizing to prevent underflow and overflow, we arrive at the dynamic broadcast and receive centralities from [14] given by

$$\mathrm{Db} := \mathrm{Db}^{[1]}, \qquad \mathrm{Dr} := \mathrm{Dr}^{[M]}, \tag{3}$$

where the vector sequence $\{\mathrm{Db}^{[r]}\}_{r=1}^{M+1}$ is computed iteratively by setting $\mathrm{Db}^{[M+1]} = \mathbf{1}$ and then solving

$$\left(I - \alpha A^{[r]}\right) \mathrm{Db}^{[r]} = \mathrm{Db}^{[r+1]}$$

and normalizing

$$\mathrm{Db}^{[r]} \mapsto \frac{\mathrm{Db}^{[r]}}{\| \mathrm{Db}^{[r]} \|_2},$$

for $r = M, M-1, \dots, 1$. Similarly, a vector sequence producing the receive centralities may be computed by setting $\mathrm{Db}^{[0]} = \mathbf{1}$ and then solving

$$\left(I - \alpha \left(A^{[r]}\right)^T\right) \mathrm{Db}^{[r]} = \mathrm{Db}^{[r-1]}$$

and normalizing

$$\mathrm{Db}^{[r]} \mapsto \frac{\mathrm{Db}^{[r]}}{\| \mathrm{Db}^{[r]} \|_2},$$

for $r = 1, 2, \dots, M$. Here $A^T$ denotes the transpose of $A$.

## 4 Experimental Results

### 4.1 Comparison of Network Centrality Measures

Using the holiday travel based active node network sequence described in section 2, we now compare the six centrality measures outlined in section 3. In order to apply the measures designed for static networks, we formed a single thresholded binarized network, $B$. To do this, we first formed the time-aggregate matrix $A_{\text{sum}} := \sum_{k=1}^{M} A^{[k]}$. Then we thresholded based on a value $\theta$, so that

$$(B)_{ij} = \begin{cases} 1 & \text{if } (A_{\text{sum}})_{ij} \geq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\theta$ is chosen so that the number of edges in $B$ matches, as closely as possible, the average number of edges in $\{A^{[k]}\}_{k=1}^{M}$. For convenience, we use the following descriptors:

- **Katz broadcast** and **Katz receive** denote the centrality measures in (1) applied to the thresholded binarized network. We used $\alpha = 0.9/\rho(B)$.
- **Dynamic broadcast** and **dynamic receive** denote the centrality measures (3) on the active node subnetwork sequence. We used $\alpha = 0.9/\max_k \rho(A^{[k]})$.
- **Out degree** and **in degree** denote the row sums and column sums of $A_{\text{sum}}$ respectively; the rankings based on these measures are equivalent to the $\alpha \to 0$ rankings from dynamic broadcast and receive.

Because our aim is to indentify influential Tweeters, we intuitively expect the three broadcast-based measures (out degree, Katz broadcast and dynamic broadcast) to be more useful than the three receive-based measures (in degree, Katz receive and dynamic receive) in this context.

Each of these six centrality measures produces a vector in $\mathbb{R}^{590}$, which can be used to determine (up to ties) a ranking, that is, a permutation of the integers 1 to 590. There are, of course, many ways to compare these different measures. The upper panel in Table 1 shows the Kendall tau and Spearman rho correlation coefficients for each pairwise combination of measures. In the context of using the measures to identify important nodes, rather than looking at correlation across the entire set of centralities it is perhaps more meaningful to focus on those nodes that are identified as important. The lower panel in Table 1 therefore shows the overlap, that is, the number of common nodes, among the top ten and and top twenty lists in a pairwise manner. The tables indicate a slightly higher match within, rather than across, the broadcast-based meaures and the receive-based measures, although this is not completely consistent; for example Katz broadcast and Katz receive have the highest pairwise correlations.

For a visual overview, Figures 2 and 3 scatter plot the dynamic broadcast centrality against each other measure. In Figure 2 we see that dynamic broadcasting and dynamic receiving are quite different achievements. One node comes top in both measures, and from Table 1 we see that 16 nodes appear in both top 20 lists. However, the orderings within the top twenty are clearly different.

| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
|---|---|---|---|---|---|---|
| out degree | | 0.48 | 0.34 | 0.35 | 0.60 | 0.46 |
| in degree | 0.48 | | 0.43 | 0.46 | 0.47 | 0.64 |
| Katz broadcast | 0.31 | 0.42 | | 0.87 | 0.34 | 0.42 |
| Katz receive | 0.33 | 0.47 | 0.88 | | 0.36 | 0.45 |
| dynamic broadcast | 0.69 | 0.52 | 0.32 | 0.35 | | 0.49 |
| dynamic receive | 0.47 | 0.73 | 0.41 | 0.45 | 0.54 | |

| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
|---|---|---|---|---|---|---|
| out degree | | 2 | 5 | 2 | 6 | 3 |
| in degree | 6 | | 1 | 1 | 2 | 2 |
| Katz-broadcast | 11 | 3 | | 3 | 6 | 3 |
| Katz-receive | 4 | 7 | 4 | | 3 | 9 |
| dynamic broadcast | 6 | 4 | 7 | 15 | | 4 |
| dynamic receive | 4 | 5 | 5 | 18 | 16 | |

**Table 1.** Upper panel shows Kendall tau correlation across pairs of node rankings in upper triangle and Spearman rho correlation across pairs of node rankings in lower triangle. Lower panel shows overlap between top 10 across pairs of node rankings in uppper triangle and overlap between top 20 across pairs of node rankings in lower triangle

Perhaps most noticably, the fourth highest dynamic broadcaster ranks relatively poorly according to dynamic receive. Further investigation revealed that this account belongs to a travel insurance brand. The account (id = 34[2]) appears to supply automated Tweets on the subject of insurance. (In the exercise reported in subsection 4.2, the social media experts ranked this account as mid-range because the Tweets generated were not personalised according to best practice.)

In the upper left picture of Figure 3 the second highest dynamic broadcaster stands out as having a relatively low Katz broadcast measure. This account (id = 398) Tweets stories about travel. As with account 34 discussed above, there were a lot of automated Tweets. This appears to be an account that is looking to send out, rather than receive, links, and most Tweets contain links to websites—however the content of the Tweets was felt to be relevant to the topic, which is why the account appears in third place in the overall expert summary of subsection 4.2 (Table 4).

In the upper right picture of Figure 3 the first and third best Katz receivers (id = 388 and 394, respectively) are seen to have very poor dynamic broadcast measure. These accounts belong to news aggregators Tweeting about travel and other news. They passed on similar information and have a similar follower profile.

The fourth higest out degree node is seen in the lower left picture of Figure 3 to be a very poor dynamic broadcaster. This unusual account (id = 341370) Tweeted about lots of different topics but has only 35 followers. This case caused an interesting split between the social media experts during the exercise discussed in subsection 4.2. Two experts rated the account as mid range and three rated it lowest of those considered. On closer inspection, we found that the accounts which were subsequently retweeting exhibited some strange behaviour

[2] The id numbers are local to this experiment and have no further significance.

that was not obvious at first glance. Figure 4 illustrates one set of retweets, suggesting that an automated process is at work in the retweeting operation, in an effort leverage influence.

More generally, it is clear from Figure 3 and Table 1 that high out degree nodes can have very poor dynamic broadcast centrality—generating a high bandwidth does not directly translate into effective communication in this sense.

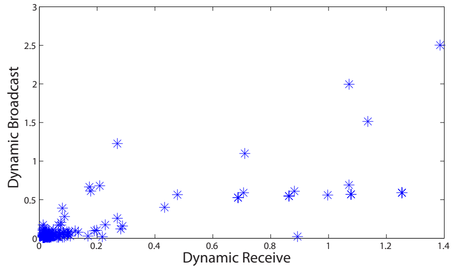

**Fig. 2.** Dynamic broadcast against dynamic receive for the active nodes.

In the lower right picture of Figure 3 there are three accounts with very high in degree that are not good dynamic broadcsters. The highest in degree account (id = 172) belongs to a holiday company based in Kauai, Hawaii, Tweeting about holidays there. The account produces some automated Tweets but they do not appear to be designed simply to publicize links. The next (id = 158) was regarded by the experts as the most heavily automated of those considered, generating Tweets on a wide range of subjects, not focused in any area, with the apparent aim of link distribution. The third (id = 31) was a news aggregator in the manner of accounts 388 and 394 discussed above.

### 4.2 Results from Social Media Experts

In order to benchmark the centrality results, we enlisted the help of five professionals working in social media who have day-to-day experience of ranking and targeting accounts based on Twitter data. It is not feasible to study by eye the full set of dynamic interaction data across the 590 active nodes—indeed, this is a key motivation for the use of automated tools. Hence, in collaboration with social media professionals, and with the aid of the six centrality measures, we focused attention on a list of 41 accounts that were felt to be highly relevant. The five experts were then given access to the full details of the Tweets from this list, including the content of their messages, and asked to rank them in order of importance. They had no knowledge of the six centrality rankings.

Table 2 records the level of consistency between the five experts, in terms of Kendall tau correlations across the 41 accounts and overlap between the top 10 in each list. We see that although the correlation is generally positive, there

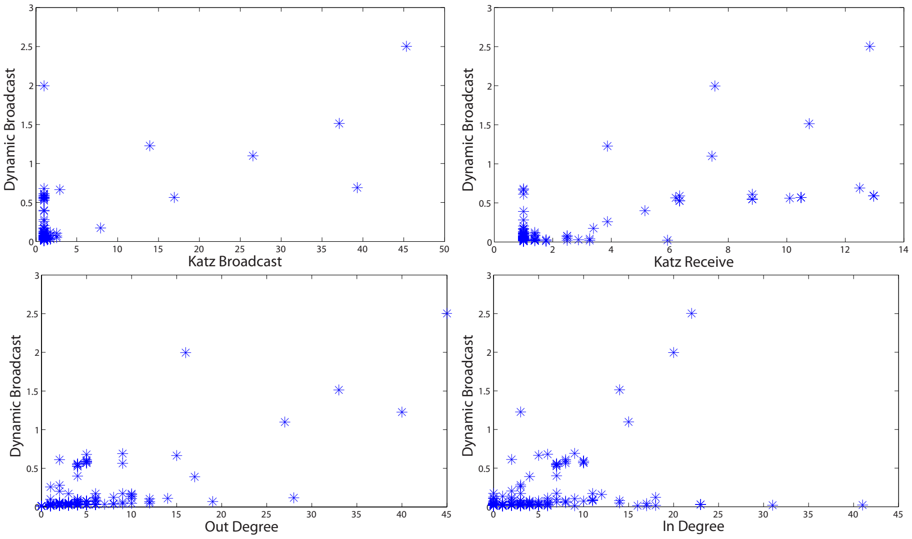

**Fig. 3.** Dynamic broadcast against: upper left: Katz broadcast, upper right: Katz receive, lower left: out degree, lower right: in degree, for the active nodes.

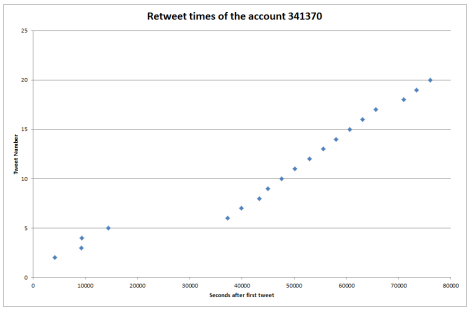

**Fig. 4.** Retweet times for a Tweet emerging from account id 341370.

is some considerable variation between the views. Hence, although we regard this information as providing a very useful guide, we do not see it as a “gold standard” with which to judge centrality measures in this context.

| | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 |
|---|---|---|---|---|---|
| Expert 1 | | -0.10 | 0.93 | 0.19 | 0.33 |
| Expert 2 | 5 | | -0.10 | 0.31 | 0.14 |
| Expert 3 | 10 | 3 | | 0.20 | 0.37 |
| Expert 4 | 6 | 5 | 6 | | 0.55 |
| Expert 5 | 6 | 5 | 6 | 5 | |

**Table 2.** Upper: Kendall tau correlation between rankings of the 41 Tweeters from pairs of experts. Lower: overlap amongst top ten in rankings of the 41 Tweeters from pairs of experts.

For Table 3 we merged the five different expert rankings of the 41 nodes, giving equal weight to each, into a single list. We then compared this ‘average expert’ with the rankings of these 41 nodes produced by each of the six centrality measures. We show the top ten overlap. Comparing with the results in Table 2, it may be argued that at least three of the centrality measures are almost indistinsguishable from experts in this sense. To give more insight, Table 4 shows the top 10 list for the averaged expert and the three broadcast-based centralities. We see that dynamic broadcast has a top three that includes two of the experts’ top three. Out degree and Katz broadcast have one such ‘correct’ answer in their top three. We also note that the centrality rankings are closer to each other than to the average expert, in terms of overlap.

| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
|---|---|---|---|---|---|---|
| Overlap | 4 | 3 | 2 | 1 | 3 | 2 |

**Table 3.** Overlap amongst top ten for each of the six centrality meaures against the average over five experts.

| average expert | out degree | Katz broadcast | dynamic broadcast |
|---|---|---|---|
| 397 | 74 | 74 | 74 |
| 362 | 34 | 302 | 398 |
| 398 | 362 | 362 | 362 |
| 341 | 341370 | 358 | 34 |
| 289 | 358 | 375 | 358 |
| 345 | 71 | 34 | 302 |
| 462 | 345 | 341 | 397 |
| 212 | 398 | 352 | 352 |
| 71 | 352 | 200 | 373 |
| 18 | 484 | 409 | 380 |

**Table 4.** Account ids in rank order from 1 to 10. Column 1: average over five experts. Column 2: out degree. Column 3: Katz broadcast. Column 4: dynamic broadcast.

## 5 Summary and Future Work

Our aim in this work was to investigate the use of network centrality measures on appropriatelty processed Twitter data as a means to target influential nodes. We found that these measures can extract value, both in isolation and when combined, especially when the time-dependent nature of the interactions is incorporated. In particular, benchmarking against the views of five experts in social media showed that the dynamic broadcast centrality results are, in the sense of overlap at the important upper end, hard to distinguish from hand curated expert rankings.

There are many open questions and remaining challenges in this area. Obvious issues include the best way to choose algorithmic parameters, such as the time window size, $\Delta t$, and the Katz downweighting parameter, $\alpha$. For long time periods, or real-time monitoring, it would also be of interest to consider downweighting information over time, as described in [13]. A bigger challenge is detecting, categorizing and dealing with accounts that generate automated Tweets. Here, it may be preferable to leave the elegant but simplified network viewpoint and dig down into the precise correlations over time of account activity.

**Acknowledgements** will appear in the de-anonymized version.

## References

1. P. Bajardi, A. Barrat, F. Natale, L. Savini, and V. Colizza, *Dynamical patterns of cattle trade movements*, PLoS ONE, 6 (2011), p. e19869.
2. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, *Everyone's an influencer: quantifying influence on Twitter*, in Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, New York, NY, USA, 2011, ACM, pp. 65–74.
3. A.-L. Barabási, *The origin of bursts and heavy tails in human dynamics*, Nature, 435 (2005), pp. 207–211.
4. D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton, *Dynamic reconfiguration of human brain networks during learning*, Proc. Nat. Acad. Sci., 108 (2011), p. doi: 10.1073/pnas.1018985108.
5. K. Berman, *Vulnerability of scheduled networks and a generalization of Menger's Theorem*, Networks, 28 (1996), pp. 125–134.
6. F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes, *Social network analysis and mining for business applications*, ACM Trans. Intell. Syst. Technol., 2 (2011), pp. 22:1–22:37.
7. S. P. Borgatti, *Centrality and network flow*, Social Networks, 27 (2005), pp. 55–71.
8. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, *Measuring user influence in Twitter: The million follower fallacy*, in in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social, 2010.
9. N. Eagle, A. S. Pentland, and D. Lazer, *Inferring friendship network structure by using mobile phone data*, Proc. Nat. Acad. Sci., 106 (2009), pp. 15274–15278.

10. E. Estrada, *The Structure of Complex Networks*, Oxford University Press, Oxford, 2011.
11. E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith, *A conceptual and operational definition of 'social role' in online community*, in Proceedings of the 42nd Hawaii International Conference on System Sciences, Los Alamitos, CA, USA, 2009, IEEE Computer Society, pp. 1–11.
12. P. Grindrod and D. J. Higham, *Evolving graphs: Dynamical models, inverse problems and propagation*, Proc. Roy. Soc. A, 466 (2010), pp. 753–770.
13. P. Grindrod and D. J. Higham, *A matrix iteration for dynamic network summaries*, SIAM Review, (2012, to appear).
14. P. Grindrod, D. J. Higham, M. C. Parsons, and E. Estrada, *Communicability across evolving networks*, Physical Review E, 83 (2011), p. 046120.
15. P. Holme, *Network reachability of real-world contact sequences*, Physical Review E, 71 (2005).
16. D. Huffaker, *Dimensions of leadership and social influence in online communities*, Human Communication Research, 36 (2010), pp. 593–617.
17. L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Rav, C. Rizzo, and A. E. Tozzi, *Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors*, PLoS ONE, 6 (2011), p. e17144.
18. L. Katz, *A new index derived from sociometric data analysis*, Psychometrika, 18 (1953), pp. 39–43.
19. D. Kempe, J. Kleinberg, and A. Kumar, *Connectivity and inference problems for temporal networks*, J. Comput. Syst. Sci., 64 (2002), pp. 820–842.
20. H. Kim, J. Tang, R. Anderson, and C. Mascolo, *Centrality prediction in dynamic human contact networks*, Comput. Netw., 56 (2012), pp. 983–996.
21. G. Kossinets, J. Kleinberg, and D. Watts, *The structure of information pathways in a social communication network*, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge Discovery and Datamining, KDD '08, New York, NY, USA, 2008, ACM, pp. 435–443.
22. H. Kwak, C. Lee, H. Park, and S. Moon, *What is Twitter, a social network or a news media?*, in Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA, 2010, ACM, pp. 591–600.
23. K. Lerman, R. Ghosh, and T. Surachawala, *Social contagion: An empirical study of information spread on digg and twitter follower graphs*, CoRR, abs/1202.3162 (2012).
24. A. V. Mantzaris and D. J. Higham, *A model for dynamic communicators*, European Journal of Applied Mathematics, to appear (2012).
25. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, *Community structure in time-dependent, multiscale, and multiplex networks*, Science, 328 (2010), pp. 876–878.
26. M. Newman, *A measure of betweenness centrality based on random walks*, Social Networks, 27 (2005), pp. 39–54.
27. M. E. J. Newman, *Networks an Introduction*, Oxford Univerity Press, Oxford, 2010.
28. D. A. Shamma, L. Kennedy, and E. F. Churchill, *In the limelight over time: Temporalities of network centrality*, Proceedings of the 29th international conference on Human factors in computing systems CSCW 2011, ACM, 2011.
29. J. Tang, S. Scellato, M. Musolesi, C. Mascolo, and V. Latora, *Small-world behavior in time-varying graphs*, Phys. Rev. E, 81 (2010), p. 05510.