Soc. Netw. Anal. Min. (2013) 3:1311–1323
DOI 10.1007/s13278-013-0143-7

ORIGINAL ARTICLE

Discovering and validating influence in a dynamic online social network

Peter Laflin · Alexander V. Mantzaris · Fiona Ainley · Amanda Otley · Peter Grindrod · Desmond J. Higham

Received: 12 March 2013 / Revised: 28 August 2013 / Accepted: 3 October 2013 / Published online: 19 October 2013

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 commercial 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), without requiring further information about the account type or message content, 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, given access to the full tweet content, who judge twitter accounts as part of their professional duties. We find that the dynamic centrality measures add value to the expert view, and can be hard to distinguish from an expert in terms of who they place in the top ten. These algorithms, which involve sparse matrix linear system solves with sparsity driven by the underlying network structure, can be applied to very large-scale networks. We also test an extension of the dynamic centrality measure that allows us to monitor the change in ranking, as a function of time, of the twitter accounts that were eventually deemed influential.

P. Laflin · F. Ainley · A. Otley
Bloom Agency, Green Sand Foundry, Marshalls Mills, Marshall Street, Leeds LS11 9YJ, UK

A. V. Mantzaris · D. J. Higham (✉)
Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK
e-mail: d.j.higham@strath.ac.uk

P. Grindrod
Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK

1 Motivation

Centrality measures have proved to be extremely useful for identifying significant players in an interaction network (Newman 2010; Wasserman and Faust 1994). 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 (Eagle et al. 2009; Grindrod et al. 2011), email activity (Barabási 2005; Grindrod et al. 2011), online social interaction (Boutet et al. 2013; Cazabet et al. 2012; Kashoob and Caverlee 2012; Tang et al. 2010), geographical proximity of mobile device users (Isella et al. 2011), voting and trading patterns (Bajardi et al. 2011; Mucha et al. 2010) and neural activity (Bassett et al. 2011; Grindrod 2010).

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. To motivate the study, we assume that the aim is to find suitable targets from a marketing perspective; that is, to identify individuals with whom it would be timely and fruitful to build relationships, with a view to propagating information to other relevant parties in the network. However, we note that these techniques also have applications in many other fields, including health, politics and security, where it is desirable to know the optimal places to disseminate information or monitor activity. In this social interaction setting, the idea of key players, who influence the actions of others, is intuitively reasonable. Empirical evidence is given in Eric et al. (2009) for discussion catalysts in an online community who are "responsible for the majority of messages that initiate long threads." Further, Huffaker (2010) identifies online 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 Mantzaris and Higham (2012) 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 on 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 knock-on effect of an interaction must be quantified if key players are to be identified. A recent business-oriented survey (Bonchi et al. 2011; Sect. 4] lists network dynamics as a key technical challenge, and the authors in Shamma et al. (2011) argue that "the temporal aspects of centrality are underrepresented."

Several recent articles have addressed the issue of discovering important or influential players in networks derived from Twitter data. The work in Eytan et al. (2011) 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 Kristina et al. (2012) looked at large-scale information spread on the Twitter follower graph to measure global activity. The authors in Meeyoung et al. (2010) studied a large-scale Twitter follower graph and compared three measures 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, Haewoon et al. (2010) 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 Meeyoung et al. (2010) and Haewoon et al. (2010) 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

- focussing on subject-specific tweets of interest in a typical marketing application,
- building the interactions between tweeters on this topic and recording them in a novel, time-stamped form that we call the active node subnetwork sequence,
- 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 full tweet contents, and
- quantifying and visualizing the time-dependent trajectories of the influence scores of the "big hitters."

2 Building the active node subnetwork sequence

Twitter is a means to send out information over a well-defined tweeter–follower network. This brings to life a scenario that social scientists have for many years been using as a theoretical device to develop concepts and measures. 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 (Borgatti 2005; Estrada 2011; Newman 2010). In this setting, most centrality measures are defined through, or can be reinterpreted from, the notion of studying random walks along the edges (Newman 2005), or exploiting the combinatorics of geodesics, paths, trails or walks (Borgatti 2005). 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 the use of this type of data

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 exchange, 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.
Link aggregation: if A and B meet only today, and B and C meet only tomorrow, the (undirected) aggregate network for the week contains the path from C to B to A, even though this traversal through the network violates time's arrow. Hence, a static aggregation leads to systematic overestimation of the spread of information (or disease) around a dynamic network (Holme and Saramäki 2012; Mantzaris and Higham 2013; Rocha et al. 2011).

By contrast, Twitter data allows us to sidestep, to some extent, 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 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.
- *Link aggregation*: at the cost of storage and computational effort, we can avoid simple aggregation and perform analysis on the fine-grained, time-stamped interaction data.

Twitter's follower graph, where nodes represent users and a directed link connects a user to a follower, has been studied, for example, in Meeyoung et al. (2010), Haewoon et al. (2010) and Kristina et al. (2012). 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. This pruning exercise generally has the effect of reducing the size of the network considerably, which is important if we wish to consider global tweets about popular topics over long time scales, where data sets become large.

To be precise, we use the Twitter feed to construct an active node subnetwork sequence in the following manner.

Definition 1 The active node subnetwork sequence may be defined as follows

- Start the clock at time t_{start}.
- Listen to all tweets that contain any of 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_{end}.
- Post-process the network by removing all nodes that have zero aggregate out degree, i.e. remove those followers 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 use the notation $t_k = t_{\text{start}} + (k-1)\Delta t$. Then, for $k = 1, 2, \ldots, M$, the kth 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).

We note that after the initial topic-based filtering, we pay no further attention to the content of the tweets. The data that we store in the time-ordered sequence of adjacency matrices is purely topological. This simplification makes the data amenable to network style centrality measures. In Sect. 4.2, we compare the results with those from social media experts who were given access to all content.

Implicit in this definition is the simplifying assumption that a tweet has an influence over a fixed period of time, Δt. It may be argued that a tweet, once sent, exists forever 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 Δt then quantifies the typical "read and respond" time.

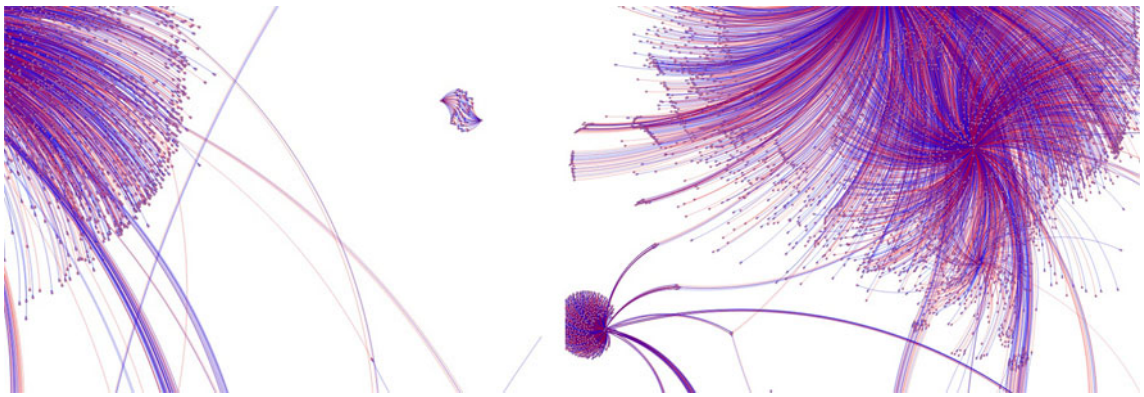


Fig. 1 Two details from the active subnode network sequence at the end of the first time window. In the left hand picture, we see an isolated community

We emphasize in particular that reducing Δt 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 Δt too large (e.g. the extreme case of one giant time window) loses 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 Δt 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 $N = 590$.

Some accounts produced extremely rapid bursts. One account tweeted 104 times in time frame 10 and a further 23 times in time frame 11. This account released a total of 127 tweets in 68 min. This motivates our decision to binarize the data within each window—in this way we do not keep track of how many times an account tweets, but rather we represent the fact they did tweet in that time frame. This prevents the overall result from being influenced by accounts using a high volume of automated tweets. The 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, Fig. 1 visualizes two portions of the the network at the end of the first time window. We will return to this data set in Sect. 4 when we compare centrality meaures.

3 Centrality measures

We proceed by reviewing relevant concepts for a single network, represented by a binary adjacency matrix $A \in \mathbb{R}^{N\times N}$. Here, $a_{ij} = 1$ if node i has a directed link to node j and $a_{ij} = 0$ otherwise. The resolvent matrix $(I - \alpha A)^{-1}$ was proposed by Katz (1953) as a means to summarize pairwise "influence" under "attenuation through intermediaries." Here the fixed parameter α governs the strength of the attenuation. For $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ denotes the spectral radius of A, that is, the maximum modulus over all the eigenvalues 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 (Estrada 2011), 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 α^p. The idea of attaching less importance to longer walks is intuitively reasonable, and Katz (1953) also points out that α may be intepreted as the chance that a message successfully traverses an edge, with probabilities being independent along each edge. It follows that the ability of nodes to broadcast and receive information can be quantified through the vector of row sums

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

$$(I - \alpha A)^{-1}\mathbf{1}, \tag{1}$$

and the vector of column sums

$$(I - \alpha A^{\mathrm{T}})^{-1}\mathbf{1}, \tag{2}$$

respectively. Here $\mathbf{1} \in \mathbb{R}^N$ is the vector with all entries equal to one, and the ith components of the vectors in (1) and (2) relate to node i. We use A^{T} to denote the transpose of A. Rather than inverting $I - \alpha A$ and $I - \alpha A^{\mathrm{T}}$, 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

$$(I - \alpha A)\mathrm{Kb} = \mathbf{1}, \qquad (I - \alpha A^{\mathrm{T}})\mathrm{Kr} = \mathbf{1}. \tag{3}$$

In this case the ith 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, with a larger value giving a higher ranking.

In the limit of extreme attenuation, $\alpha \to 0$, longer walks make a negligible contribution in (3), and ignoring uniform shifts and scalings that do not alter the rankings, the measures Kb and Kr collapse to out degree and in degree, respectively, that is,

$$(\deg_{\mathrm{out}})_i = \sum_{j=1}^{N} a_{ij}, \qquad \text{and} \qquad (\deg_{\mathrm{in}})_j = \sum_{i=1}^{N} a_{ij}. \tag{4}$$

Of course, these two quantities are also widely used as centrality measures in their own right (Estrada 2011; Newman 2010).

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 (Petter 2005; Kim et al. 2012; Gueorgi et al. 2008; Mucha et al. 2010). We focus here on the dynamic walk notion from Grindrod et al. (2011), which produces generalizations of the Katz centrality measures (3) that are feasible for large-scale network computations. Using the notation introduced in Sect. 2, where $(A^{[k]})_{ij} = 1$ denotes the presence of an edge between nodes i and j at time t_k, the following definition was made in Grindrod et al. (2011).

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

Dynamic walks are easily counted by forming appropriate matrix products. 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}.

Exploiting the idea of Katz in this time-dependent setting, we may count dynamic walks in such a way that walks of length w are downweighted by α^w. This leads to the expression

$$\left(I - \alpha A^{[1]}\right)^{-1}\left(I - \alpha A^{[2]}\right)^{-1} \cdots \left(I - \alpha A^{[M]}\right)^{-1} \tag{5}$$

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

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 (Grindrod et al. 2011) given by

$$\mathrm{Db} := \mathrm{Db}^{[1]}, \qquad \mathrm{Dr} := \mathrm{Dr}^{[M]}, \tag{6}$$

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 $r = M, M-1, \ldots, 1$.

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

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

and normalizing

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

for $r = 1, 2, \ldots, M$.

In terms of computational cost, to compute either the broadcast or the receive centralities requires the solution of one sparse linear system per time point. In each case the matrix dimension is given by the number of nodes in the network and the level of sparsity corresponds directly to that of the interaction structure. We also note that for a given set of Twitter data, refining the time window, Δt,

creates sparser adjacency matrices; so, in this sense, the cost of the algorithm scales sublinearly with the number of time windows chosen for a given data set.

4 Experimental results

4.1 Comparison of network centrality measures

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

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

Here θ was chosen so that the number of edges in B matched, 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 (3) applied to the thresholded binarized network. We used $\alpha = 0.9/\rho(B)$, based on experience in (Grindrod et al. 2011).
- *Dynamic broadcast* and *dynamic receive* denote the centrality measures (6) 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_{sum} respectively, as in (4); the rankings based on these measures are equivalent to the rankings from dynamic broadcast and receive in the limit of vanishing attenuation parameter, $\alpha \to 0$.

Because our aim is to identify 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 of the network nodes, that is, a permutation of the integers 1–590. There are, of course, many ways to compare these different measures. The upper panel in Table 1 shows the Kendall tau and Spearman ρ 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 arguably 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, Figs. 2, 3, 4, 5 and 7 scatter plot the dynamic broadcast centrality against each other measure. In Fig. 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 twenty lists. However, the orderings within the top twenty are clearly different.

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 and lower panel shows overlap between top ten across pairs of node rankings in uppper triangle and overlap between top twenty across pairs of node rankings in lower triangle

	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		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	

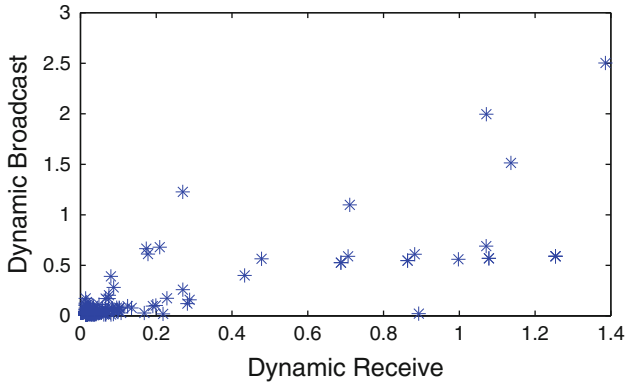


Fig. 2 Dynamic broadcast against dynamic receive for the active nodes

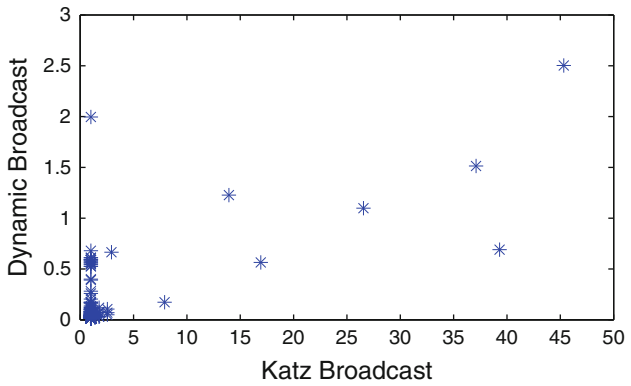


Fig. 3 Dynamic broadcast against Katz broadcast for the active nodes

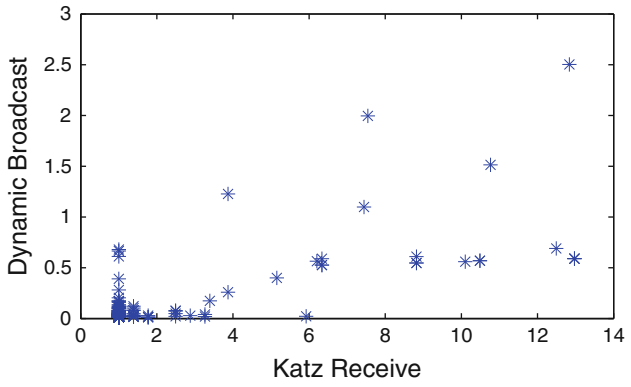


Fig. 4 Dynamic broadcast against Katz receive for the active nodes

Perhaps most noticeably, 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

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

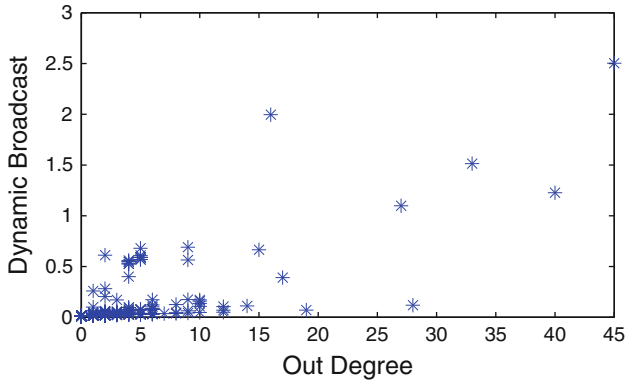


Fig. 5 Dynamic broadcast against out degree for the active nodes

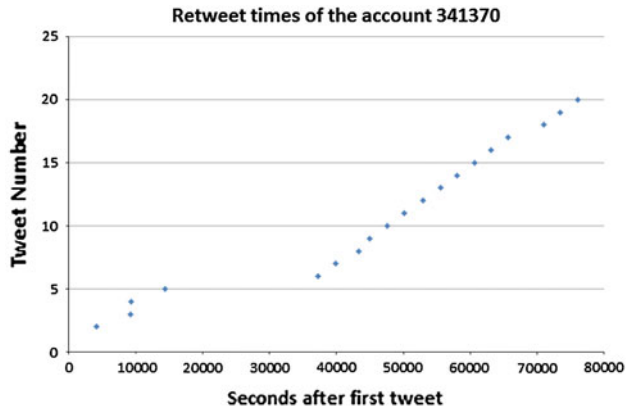


Fig. 6 Retweet times for a tweet emerging from account id 341,370

insurance. (In the exercise reported in Sect. 4.2, the social media experts ranked this account as mid-range because the tweets generated were not personalized.)

In Fig. 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 highly relevant to the topic, which is why the account appears in third place in the overall expert summary of Sect. 4.2 (Table 4).

In Fig. 4 the first and third best Katz receivers (id = 388 and 394, respectively) are seen to be poor dynamic broadcasters. 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 highest out degree node is seen in the lower left picture of Fig. 5 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

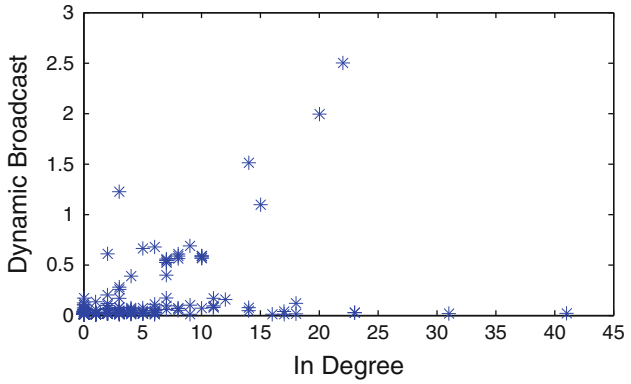


Fig. 7 Dynamic broadcast against in degree for the active nodes

the exercise discussed in Sect. 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 that was not obvious at first glance. Figure 6 illustrates one set of retweets, suggesting that an automated process is at work in the retweeting operation, in an effort to leverage influence.

More generally, it is clear from these results 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. This effect of aggregate edge counts correlating poorly with more sensitive time-respecting measures has been observed on other data sets; notably email (Mantzaris and Higham 2012; John et al. 2009).

In Fig. 7 there are three accounts with very high in degree that are not good dynamic broadcasters. 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

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. These experts look for influential nodes in the network as a means for targeted intervention and engagement. 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, based on the automated ranking results and further account-level information. 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 Upper Kendall tau correlation between rankings of the 41 tweeters from pairs of experts and lower overlap amongst top ten in rankings of the 41 tweeters from pairs of experts

	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 records the level of consistency between the five experts, in terms of Kendall tau correlations across the 41 accounts and overlap between the top ten in each list. We see that although the correlation is generally positive, there is some considerable variation between the views. Hence, although we regard this information as providing a very useful guide, we do not present it as a "gold standard" with which to judge centrality measures in this context.

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 'averaged 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 indistinguishable from experts in this sense. We also see that within each of the three centrality categories, degree, Katz and dynamic, the broadcast version performed better than receive. This agrees with our initial intuition that, in the context of targeting influential users whose behaviour affects others in the network, information dispersal is more relevant than information gathering. To give more detail, Table 4 shows the top ten 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.

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

	Out degree	In degree	Katz broadcast	Katz receive	Dynamic broadcast	Dynamic receive
Overlap	4	3	2	1	3	2

Table 4 Account ids in rank order from 1 to 10

Averaged 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

Column 1: average over five experts. Column 2: out degree. Column 3: Katz broadcast. Column 4: dynamic broadcast

5 Influence trajectories

The product (5) is designed for the scenario where data is supplied over a well-defined time period $[t_1,t_M]$ and we we wish to summarize the communicability at the final time point, t_M. This is appropriate in the setting of Sect. 4, where we wish to measure influence to drive some sort of intervention at time t_M. However, there are other, related, applications where we are interested in monitoring the levels of influence over an unspecifed time period; for example, we may have the freedom to wait for an 'optimal' time to intervene in the network, or we may wish to compute some overall running summary of the influence scores of certain players of interest. Further, due to its underlying combinatorical nature, the product (5) gives equal weight to all walks of equal length that took place in the time period $[t_0,t_M]$ and, by construction, the dynamic centrality cannot decrease if we add later time points. If we are interested in real-time monitoring over a long time period, or predicting future behaviour, then it is more natural to focus on current and recent activity—dynamic walks should be downweighted according to their temporal age, on the grounds that messages lose their relevance over time. Based on this reasoning, in (Grindrod and Higham 2013) an extension to the dynamic broadcast and receive measures was given where in addition to the walk-length attenuation parameter α, we have a temporal parameter $b > 0$ which filters out old activity. More precisely, each walk is downweighted by the age-dependent factor e^{-bt}, where t denotes the time that has elapsed since the walk began.

The iteration proposed and tested in (Grindrod and Higham 2013) takes the form

$$\mathcal{S}^{[k]} = (I + \mathrm{e}^{-b\Delta t}\mathcal{S}^{[k-1]})(I - \alpha A^{[k]})^{-1} - I, \quad k = 1, 2, 3\ldots, \tag{7}$$

where, for convenience, $\mathcal{S}^{[0]} = 0$. (We also note that the algorithm generalizes straightforwardly to the case where the time points are not equally spaced.)

To interpret the iteration (7) combinatorically, we may expand the right hand side to obtain

$$\alpha A^{[k]} + \alpha^2 {A^{[k]}}^2 + \cdots + \alpha^r {A^{[k]}}^r + \cdots \tag{8}$$

$$+e^{-b\Delta t_k}\mathcal{S}^{[k-1]} \tag{9}$$

$$+e^{-b\Delta t_k}\mathcal{S}^{[k-1]}\alpha A^{[k]} + e^{-b\Delta t_k}\mathcal{S}^{[k-1]}\alpha^2 {A^{[k]}}^2 + \cdots + e^{-b\Delta t_k}\mathcal{S}^{[k-1]}\alpha^r {A^{[k]}}^r + \cdots \tag{10}$$

We see that

- In (8) we use a purely length-weighted count for all walks that start and finish at the current time, t_k.
- In (9) counts of the "old" walks that do not involve time t_k are downweighted by the time-factor $\mathrm{e}^{-b\Delta t}$, to account for the ageing effect.
- The terms in (10) deal with newly created walks that began at an earlier time but make use of one or more edges at the current time, t_k. Temporal downweighting of $\mathrm{e}^{-b\Delta t}$ is used, because the age of each such walk has increased by Δt. We also include the length-down-weighting factor α^r when r new edges are used.

Overall, we see that the i, j element of the matrix $\mathcal{S}^{[k]}$ records a scaled count at time t_k of the number of dynamic walks from i to j that can be taken with the time-ordered sequence $A^{[1]}, A^{[2]}, A^{[3]}, \ldots, A^{[k]}$. The scaling takes account of the length and age of each walk via the product of

- a factor α^w for walks of length w, and
- a factor e^{-bt} for walks that began t time units ago.

The iteration (7) involves sparse linear system solves, and hence is applicable to large-scale networks. We may obtain node-based broadcast and receive scores by aggregating the ability of a node to communicate with, or receive communication from, every other node. In this way,

$$\mathcal{S}^{[k]}\mathbf{1} \quad \text{and} \quad {\mathcal{S}^{[k]}}^{\mathrm{T}}\mathbf{1}, \tag{11}$$

give running versions of the dynamic broadcast and receive communicabilities Db and Dr.

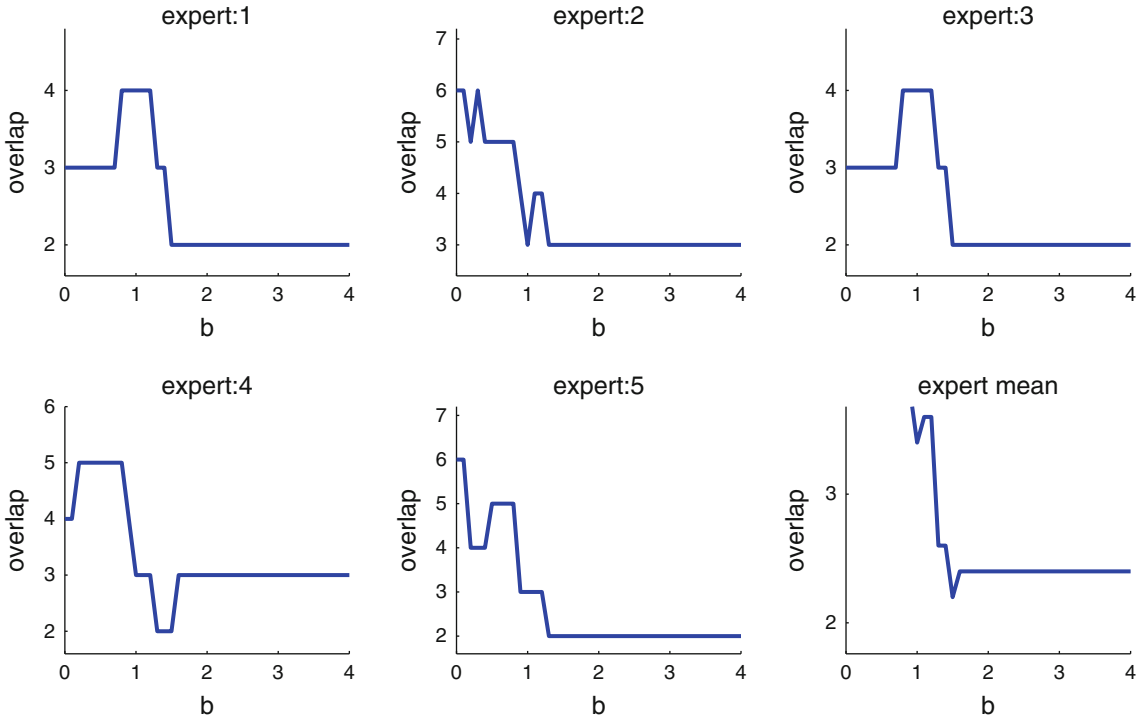


Fig. 8 Top ten overlaps over 41 key nodes, as a function of time downweighting parameter, between dynamic broadcast ranking, based on (11), and the views of each individual social media expert. Averaged expert is used for the *bottom right picture*

The temporal parameter, b, allows us to interpolate between two extremes

- $b = 0$ (no downscaling in time) reproduces the dynamic broadcast and receive rankings used in Sect. 4,
- $b \to \infty$, so that $\mathrm{e}^{-b\Delta t_k} \equiv 0$, (complete downscaling in time) reproduces the Katz static centrality applied to the single adjacency matrix at the current time point, ignoring all earlier activity.

Our aim is now to compute this running summary on the active node subset from Sect. 4. We will perform two types of tests

1. examine whether the increased flexibilty afforded by the parameter b can improve the influence rankings relative to the social media experts,
2. plot the 'influence trajectory' over time of the Twitter accounts that were eventually ranked highest.

In each test we used the 590 active node subset with the same value for the length-based downweighting parameter, α, as for the previous tests. In the first test we ran our experiment over a range of β values between zero and four. We nominally set $\Delta t = 1$, so that $\beta = 4$ corresponds to downweighting walks that started s time steps ago by a factor e^{-4s}; this means, for example, that walks aged 5 time steps are hugely discounted by the factor $e^{-20} \approx 2 \times 10^{-9}$.

In Fig. 8, we show the overlap among the 41 nodes supplied to the social media experts between (a) the top ten based on final time broadcast ranking in (11) and (b) the top ten based on the views of each individual social media expert. We also show, in the bottom right corner, the same comparison against the averaged expert ranking. We see that the overlap generally decreases with increasing b, and that values of b close to zero, corresponding to very little downweighting over time, give the best results. This is consistent with the fact the social media experts were given the Twitter data as a single entity, with no suggestion that more recent exchanges should be given priority.

In Figs. 9 and 10 instead of overlap we show how the Kendall tau and Spearman rho correlation coefficients vary in terms of b. In these cases, of course, we are measuring across all 41 Twitter accounts, not just those deemed to be the most important. From this perspective there is little evidence for any particular choice of b being optimal.

In the context of this Twitter study, where we have generated data over a prespecified time period, it is perhaps more useful to regard the dynamic rankings (11) as a means to visualize the influence trajectories over time. In Fig. 11 we therefore show the ranking, out of the full set of 590, of the nodes that eventually became the top five. We show these results for $b = 0$, 0.2, 0.5 and 2. (In this case there was strong consistency between final the top five for the cases $b = 0$, 0.2, 0.5, with four nodes appearing in each

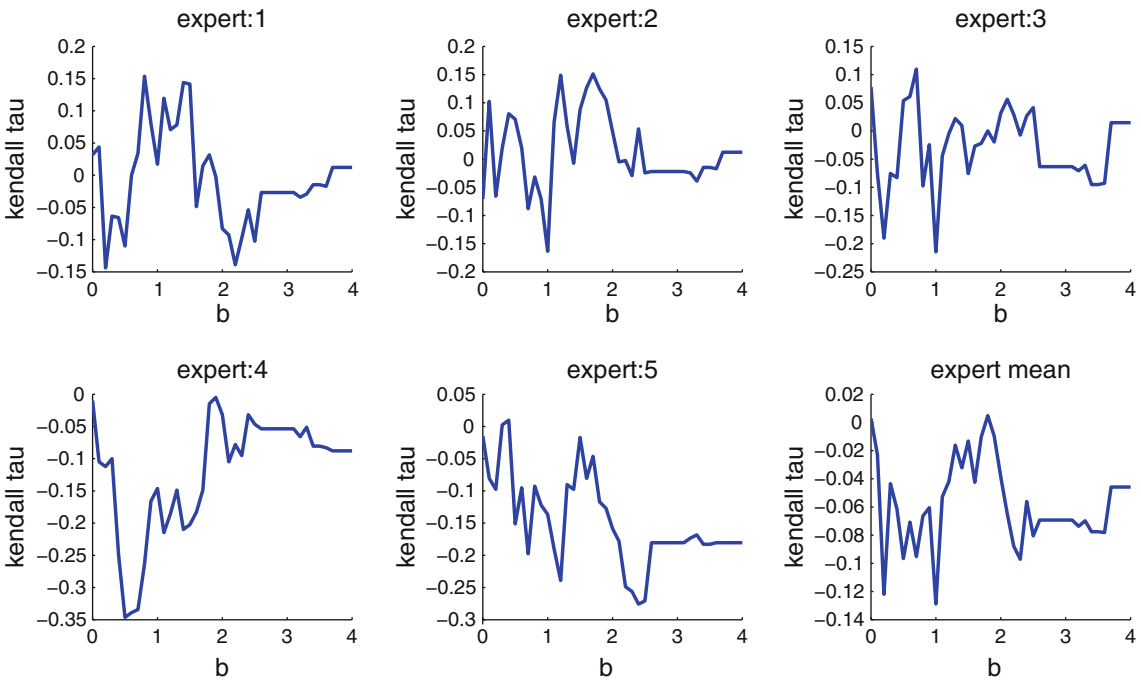


Fig. 9 Kendall tau correlation for the full set of 590 nodes, as a function of time downweighting parameter, between dynamic broadcast ranking, based on (11) and the views of each individual social media expert. Averaged expert is used for the *bottom right picture*

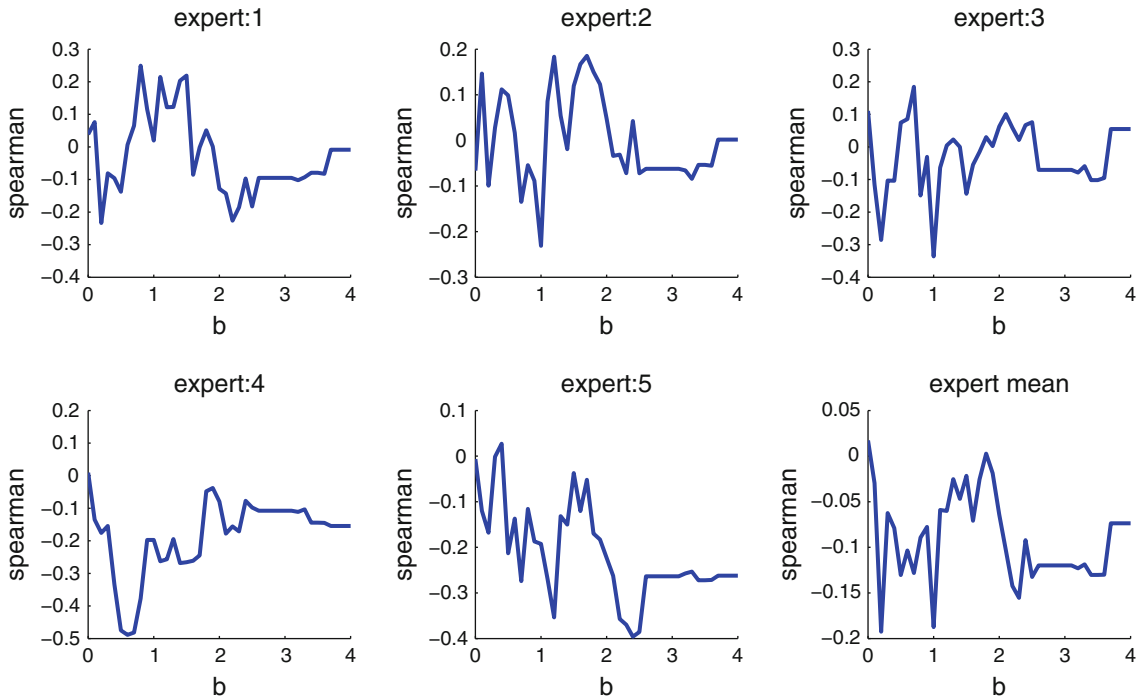
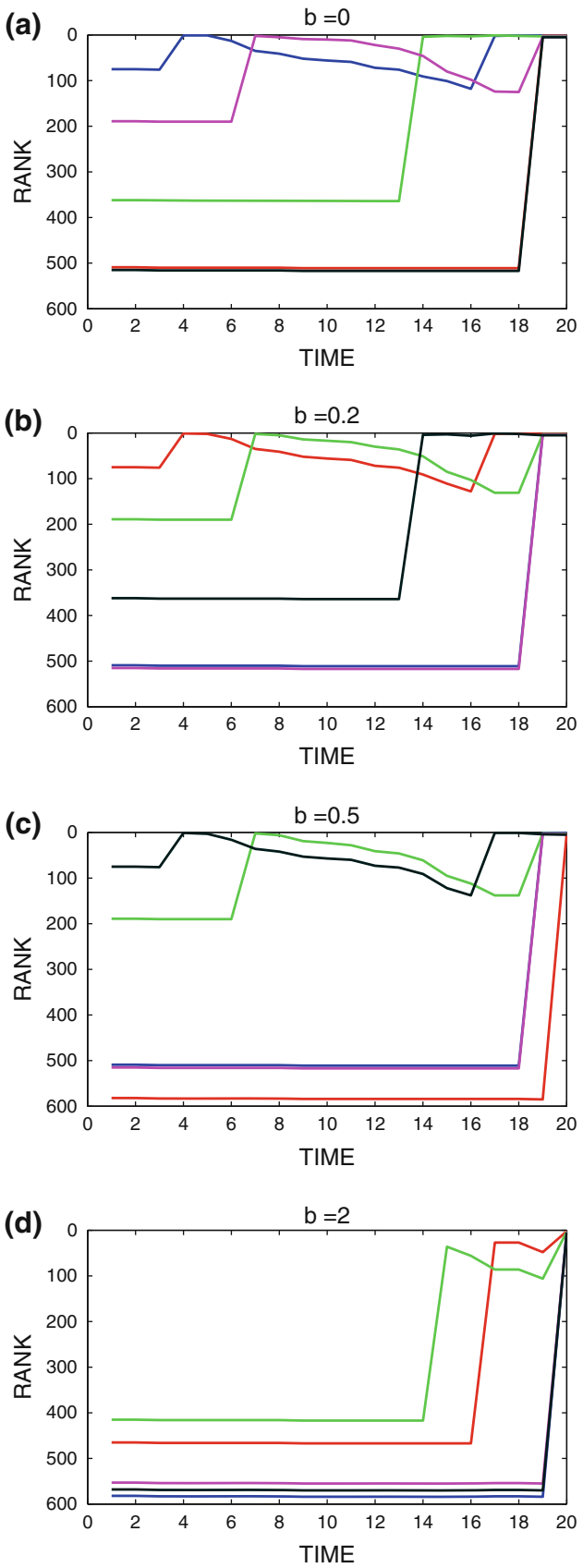


Fig. 10 Spearman rho correlation for the full set of 590 nodes, as a function of time downweighting parameter, between dynamic broadcast ranking, based on (11) and the views of each individual social media expert. Averaged expert is used for the *bottom right picture*

list, but the $b = 2$ case produced a completely distinct top five). If we regard the final time point as our key target date, we see from Fig. 11 that the choices of b equal to 0, 0.2 and 0.5 identify their key nodes more quickly (at around the 15th time window) than the more snapshot-based $b = 2$ version that gives preference to more recent data. These conclusions agree with other tests in (Grindrod and Higham 2013) that focused on centrality prediction



◀ **Fig. 11** Journey to the top of the rankings: for $b = 0$, 0.2, 0.5 and 2 we show the dynamic rankings (11) at each time t_k for the nodes that eventually became the top five out of the 590. *Colours* indicate final ranking: 1st *blue*, 2nd *red*, 3rd *green*, 4th *magenta*, 5th *black*

with email data, where making some use of historical interaction, rather than focusing entirely on current activity, was found to be effective in terms of improving predictions.

6 Summary and future work

Our aim in this work was to investigate the use of network centrality measures on appropriately processed Twitter data as a means to target influential nodes. To do this, we took the novel step of defining the active node subnetwork sequence, which gives a time-varying summary of the dynamic interactions. We then applied recently derived dynamic centrality measures for the first time in this type of study. 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. Of course, the computional algorithms, which use only toplogical information and ignore tweet content, can be applied to much larger scale data than that amenable to human curation. At the very least, computing the automated rankings could be viewed as a useful pre-processing and filtering step before experts become involved.

We also considered the use of a recently proposed running-summary iteration that places less emphasis on historical activity. This extension of the dynamic centrality measure was shown to have potential for summarizing and visualizing the rankings as a function of time, a topic which is useful for long-time monitoring and prediction purposes.

There are many open questions in this area. Remaining issues include the best way to choose algorithmic parameters, such as the time window size, Δt, Katz down-weighting parameter, α, and time decay parameter, b. 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.

In future work we will derive equivalent centrality measures valid for continuous time, avoiding the need for time step selection.

Acknowledgments Alexander V. Mantzaris, Desmond J. Higham and Peter Grindrod thank the EPSRC and RCUK Digital Economy programme for support through the project Mathematics of Large Technological Evolving Networks (MOLTEN). Desmond J. Higham was also supported by a Royal Society Wolfson Award and a Leverhulme/Royal Society Senior Fellowship. Peter Laflin, Fiona Ainley and Amanda Otley thank the Technology Strategy Board of the UK for funding the SMART project entitled Digital Business Analytics for Decision Makers. Work done on that project has contributed to the knowledge shared in this paper, especially with regard to building networks from the data. They also thank colleagues at Bloom Agency for allowing them time to work on this project, outside of their usual client workload. We thank Alex Craven, Phil Jefferies and Claire Hunter-Smith for liaising with social media experts and coordinating their feedback and comments. An earlier version of this document appeared in the Proceedings of Social Informatics 2012, Lausanne.

References

Bajardi P, Barrat A, Natale F, Savini L, Colizza V (2011) Dynamical patterns of cattle trade movements. PLoS One 6:e19869

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. in: Proceedings of the fourth ACM international conference on web search and data mining, WSDM '11, ACM, New York, pp 65–74

Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435:207–211

Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST (2011) Dynamic reconfiguration of human brain networks during learning. Proc Nat Acad Sci. doi:10.1073/pnas.1018985108

Bonchi F, Castillo C, Gionis A, Jaimes A (2011) Social network analysis and mining for business applications. ACM Trans Intell Syst Technol 2:22:1–22:37

Borgatti SP (2005) Centrality and network flow. Soc Netw 27:55–71

Boutet A, Kim H, Yoneki E (2013) What's in Twitter, I know what parties are popular and who you are supporting now! Soc Netw Anal Min. doi:10.1007/s13278-013-0120-1

Cazabet R, Takeda H, Hamasaki M, Amblard F (2012) Using dynamic community detection to identify trends in user-generated content. Soc Netw Anal Min 2:361–371

Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: the million follower fallacy. In: ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social

Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci 106:15274–15278

Estrada E (2011) The structure of complex networks. Oxford University Press, Oxford

Gleave E, Welser HT, Lento TM, Smith MA (2009) A conceptual and operational definition of social role in online community. In: Proceedings of the 42nd Hawaii international conference on system sciences, IEEE Computer Society, Los Alamitos, pp 1–11

Grindrod P, Higham DJ (2010) Evolving graphs: dynamical models, inverse problems and propagation. Proc R Soc A 466:753–770

Grindrod P, Higham DJ (2013) A matrix iteration for dynamic network summaries. SIAM Rev 55:118–128

Grindrod P, Higham DJ, Parsons MC, Estrada E (2011) Communicability across evolving networks. Phys Rev E 83:046120

Holme P (2005) Network reachability of real-world contact sequences. Phys Rev E 71:046119

Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519:97–125

Huffaker D (2010) Dimensions of leadership and social influence in online communities. Hum Commun Res 36:593–617

Isella L, Romano M, Barrat A, Cattuto C, Colizza V, Van den Broeck W, Gesualdo F, Pandolfi E, Rav L, Rizzo C, Tozzi AE (2011) Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. PLoS One 6:e17144

Kashoob S, Caverlee J (2012) Temporal dynamics of communities in social bookmarking systems. Soc Netw Anal Min 2:387–404

Katz L (1953) A new index derived from sociometric data analysis. Psychometrika 18:39–43

Kim H, Tang J, Anderson R, Mascolo C (2012) Centrality prediction in dynamic human contact networks. Comput Netw 56:983–996

Kossinets G, Kleinberg J, Watts D (2008) The structure of information pathways in a social communication network. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and datamining, KDD '08, ACM, New York, pp 435–443

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, WWW '10, ACM, New York, pp 591–600

Lerman K, Ghosh R, Surachawala T (2012) Social contagion: an empirical study of information spread on digg and Twitter follower graphs, CoRR, abs/1202.3162

Mantzaris AV, Higham DJ (2012) A model for dynamic communicators. Eur J Appl Math 23:659–668

Mantzaris AV, Higham DJ (2013) Dynamic communicability predicts infectiousness. In: Holme P, Saramäki J (eds) Temporal networks. Springer, Berlin

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328:876–878

Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27:39–54

Newman MEJ (2010) Networks an introduction. Oxford Univerity Press, Oxford

Rocha LEC, Liljeros F, Holme P (2011) Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PloS Comput Biol 7:e1001109

Shamma DA, Kennedy L, Churchill EF (2011) In the limelight over time: temporalities of network centrality. In: Proceedings of the 29th international conference on human factors in computing systems CSCW 2011, ACM

Tang J, Musolesi M, Mascolo C, Latora V (2009) Temporal distance metrics for social network analysis. In: Proceedings of the 2nd ACM SIGCOMM workshop on online social networks (WOSN09), Barcelona

Tang J, Scellato S, Musolesi M, Mascolo C, Latora V (2010) Small-world behavior in time-varying graphs. Phys Rev E 81:05510

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge