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, Hu aker [16] identi es *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 e ect* of an interaction must be quanti ed if key players are to be identi ed. 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 in uential 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 in uenced 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 in uence: number of followers (out degree), number of retweets and number of mentions, nding 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, nding 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 in uence 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 di ers from that described above by (a) focussing on subject-speci c 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

> 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 un ltered 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-de ne 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 ows along the links and thereby to identify important actors [10, 27]. In this setting, most centrality measures are de ned 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 simpli cations that the methodology involves. For example, even if we accept that social acquaintanceship is a reasonable proxy for the links along which information ows, 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 ltering based on Tweet content, we can, in principle, record only links that are relevant to a speci c 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 prede ned 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 e ect 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.

De nition 1 *The* active node subnetwork sequence*:*

- *Start the clock at time* t_{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_{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* $t = (t_{\text{end}} - t_{\text{start}})/M$*. We will let* $t_k = t_{\text{start}} + (k-1)\,t$*. 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 de nition is the simplifying assumption that a Tweet has

On the other hand taking t

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 α governs the strength of the attenuation, and for $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ denotes the spectral radius of A, we have

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

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 α^p. The idea of attaching less importance to longer walks is intuitively reasonable, and Katz [18] also points out that α 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 ith

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* $t_{r_1} \leq t_{r_2} \leq \cdots \leq t_{r_w}$ *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 ^{w}, this leads to the expression

$$I \quad A^{[1]} \quad ^1 \quad I \quad A^{[2]} \quad ^1 \quad \cdots \quad I \quad A^{[M]} \quad ^1$$

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

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

$$\mathrm{Db} := \mathrm{Db}^{[1]}$$

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 , so that

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

Here is chosen so that the number of edges in B matches, as closely as possible, the average number of edges in $\{A^{[k]}\}_{k=1}^{M}$

	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						

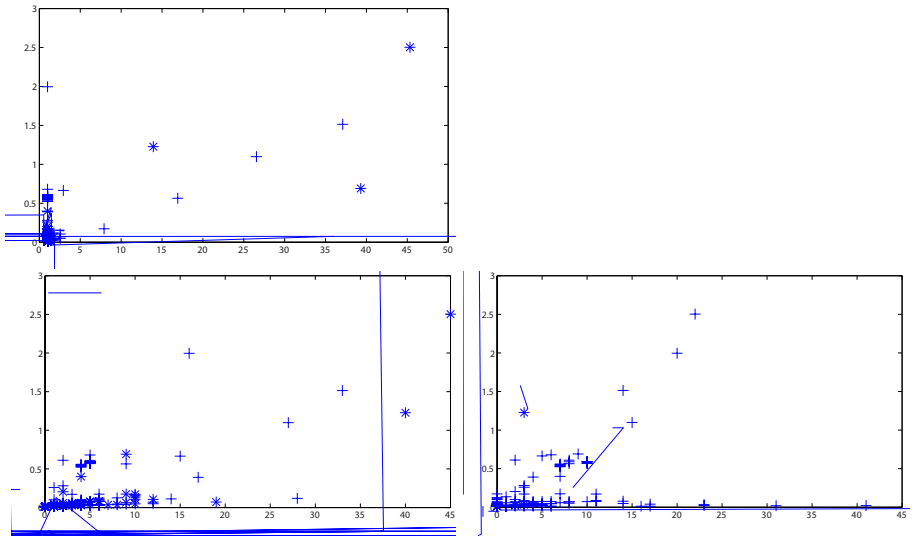

Fig. 3. Dynamic broadcast against: upper left: Katz broadcast, upper right: Katz receive, lower left: out degree, lower right: in degree, for the active nodes.

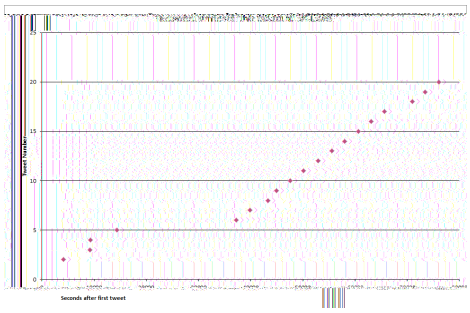

Fig. 4. Retweet times for a Tweet emerging from account id 341370.

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 ve di erent 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 that234,het cfv1.95285(rankings)ee.thada285(rankings)ee.thada285(rankings)ee.thada2

Table 2.

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 in uential 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 ve 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, t, and the Katz downweighting parameter, . 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 simpli ed 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 in uencer: quantifying in uence 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. Barabasi, *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 recon guration 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 ow*, Social Networks, 27 (2005), pp. 55{71.
8. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, *Measuring user in uence 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*