Social Network Analysis and Mining, Springer, vol. 4(1), article 204, June 2014

Modeling Individual Topic-Specific Behavior and Influence Backbone Networks in Social Media

Petko Bogdanov[†] · **Michael Busch**[†] · **Jeff Moehlis** · **Ambuj K. Singh** · **Boleslaw K. Szymanski**

[†]**These authors contributed equally.**

Received: date / Accepted: date

Abstract Information propagation in social media depends not only on the static follower structure but also on the topic-specific user behavior. Hence novel models incorporating dynamic user behavior are needed. To this end, we propose a model for individual social media users, termed a *genotype*. The genotype is a *per-topic* summary of a user's interest, activity and susceptibility to adopt new information. We demonstrate that user genotypes remain invariant within a topic by adopting them for classification of new information spread in large-scale real networks. Furthermore, we extract topic-specific influence backbone structures based on content adoption and show that their structure differs significantly from the static follower network. We also find, at the population level using a simple contagion model, that hashtags of a known topic propagate at the greatest rate on backbone networks of the same topic. When employed for influence prediction of new content spread, our genotype model and influence backbones enable more than 20% improvement, compared to purely structural features. It is also demonstrated that knowledge of user genotypes and influence backbones allows for the design of effective strategies for latency minimization of topic-specific information spread. [1]

Petko Bogdanov · Ambuj K. Singh
Department of Computer Science
University of California Santa Barbara
Santa Barbara, CA 93106
E-mail: petko@cs.ucsb.edu

Michael Busch · Jeff Moehlis
Department of Mechanical Engineering
University of California Santa Barbara
Santa Barbara, CA 93106

Boleslaw K. Szymanski
Deparment of Computer Science
Rensselaer Polytechnic Institute
110 8th St., Troy NY 12180

[1] This manuscript is an extension of the authors' earlier work presented at the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Here, the original ideas and methods are explained in further detail along with previously unpublished results.

1 Introduction

Trends and influence in social media are mediated by the individual behavior of users and organizations embedded in a follower/subscription network. The social media network structure differs from a friendship network in that users are allowed to *follow* any other user and follower links are not necessarily bi-directional. While a link enables a possible influence channel, it is not always an active entity, since a follower is not necessarily interested in all of the content that a *followee* posts. Furthermore, two individuals are likely to regard the same token of information differently. Understanding how information spreads and which links are active requires characterizing the users' individual behavior, and thus going beyond the static network structure. A natural question then arises: *Are social media users consistent in their interest and susceptibility to certain topics?*

In this work, we answer the above question by demonstrating a persistent topic-specific behavior in real-world social media. We propose a user model, termed *genotype*, that summarizes a user's topic-specific footprint in the information dissemination process, based on empirical data. The social media genotype, similar to a biological genotype, captures unique user traits and variations in different genes (topics). Within the genotype model, a node becomes an individual represented by a set of unique invariant properties.

For our particular analysis, the genotypes summarize the propensity and activity level in adoption, transformation, and propagation of information within the context of different topics. We propose a specific set of properties describing the adoption and use of topic-specific Twitter hashtags—tokens that annotate messages and allow users to participate in global discussions [1]. The model, however, applies to more general settings capturing, for example, dissemination of URLs or sentiments in the network.

We construct the genome (collection of user genotypes) of a large social media dataset from Twitter, comprised of both follower structure and associated posts. The existence of stable genotypes (behavior) leads to natural further questions: *Can this consistent user behavior be employed to categorize novel information based on its spread pattern? Can one utilize the genotypes and the topic-specific influence backbone to (i) predict likely adopters/influencers for new information from a known topic and (ii) improve the network utility by reducing latency of disseminated information?* We explore the potential of the genotype model to answer the above questions within the context of Twitter.

To validate the consistency of genotypes, we show that combining genotype-based classifiers into a composite (network-wide) classifier achieves accuracy of 87% in predicting the topic of novel hashtags that spread in the network. We extract and analyze topic-specific influence backbone networks and show that they structurally differ from the static follower network. When considering the population level dynamics, using a simple contagion model, we show that hashtags of a known topic propagate at the greatest rate on backbone networks of the same topic, and that this result is consistent with the local user model. We, then, turn to two important applications: influence prediction and topic-specific latency minimization. We achieve 20% improvement in predicting influencers/adopters for novel hashtags, based on

our model, as compared to relying solely on the follower structure. We also demonstrate that knowledge of individual user genotypes allows for effective reduction in the average time for information dissemination (a two-fold reduction by modifying the behavior of 1% of the nodes).

Our contributions, include: *(i)* proposing a genotype model for social media users' behavior that enables a rich-network analysis; *(ii)* validating the consistency of the individual genotype model; *(iii)* quantifying the differences of behavior-based influence backbones from the static network structure in a large real-world network; *(iv)* showing that the propagation rate on each topic backbone is greatest for hashtags of the same topic; and *(v)* employing genotypes and backbone structure for adopter/influencer prediction and latency minimization of information spread. Many of these results were presented in [2], yet the results and discussion of *(iv)* are entirely original to this manuscript. Furthermore, the methods and interpretations of all results herein are presented in greater detail, with emphasis on novel insights.

2 Related Work

The network structure has been central in studying influence and information dissemination in traditional social network research [3,4]. Large social media systems, different from traditional social networks, tend to exhibit relatively denser follower structure, non-homogeneous participation of nodes, and topic specialization/interest of individual users. Twitter, for example, is known to be structurally different from human social networks [5], and the intrinsic topics of circulated hashtags are central to their adoption [6].

A diverse body of research has been dedicated to understanding influence and information spread on networks, from theories in sociology [7] to epidemiology [8, 9], leading to empirical *large-scale* studies enabled by social web systems [6,10–12]. Here, we postulate that the influence structure varies across topics [13] and is further personalized for individual node pairs. Lin and colleagues [14] also focus on topic-specific diffusion by co-learning latent topics and their evolution in online communities. The diffusion that the authors of [14] predict is implicit, meaning that nodes are part of the diffusion if they use language corresponding to the latent topics. In contrast, we focus on topic-specific user genotypes and influence structures concerned with passing of observable information tokens and their temporal adoption properties.

Earlier data-centered studies have shown that sentiment [10] and local network structure [6] have an effect on the spread of ideas. The novelty of our approach is the focus on content features to which users react. Previous content-based analyses of Tweets have adopted latent topic models [15,16]. We tie both content and behavioral features to the network's individuals.

With regard to influence network structure and authoritative sources discovery, Rodriguez and colleagues [17] were able to infer the structure and dynamics of information (influence) pathways, based on the spread of memes or keywords. Bakshy et al. [18] focus on Twitter influencers who are roots of large cascades and have many followers, while Pal et al [19] adopt clustering and ranking based on structural and

content characteristics to discover authoritative users. Although the above works are similar to ours in that they focus on influence structures and user summaries, our genotype targets capturing the invariant user behavior and information spread within topics as a whole, involving a collection of topically related information parcels.

Our framework is inspired by biology and evolution, similar to Reali and Griffiths [20]. We broaden the genotype interpretation beyond word variants, and demonstrate their predictive utility. Our goal is to treat the observable content as a genetic parcel of information that users pass on to one another, while potentially introducing a delay or alteration to the message. An added benefit of this approach is that similarity of behavior toward certain types of messages among users may indicate social affinity (of interests, attitudes, etc.), provide important information about transmission paths in the network, and predict future edge formation [21].

Improving the network structure and utility has been considered in the *Influence Maximization* [3] problem with the purpose of maximizing the expected set of nodes that *eventually* adopt specific information, assuming uniform probabilistic spread to neighbors and a specific infection process. In contrast, we employ empirical user latency of information adoption and optimize for the speed of spread. This approach can be combined with our proposed node latency reduction, but we leave such extensions for future work and focus on demonstrating the utility of genotypes.

Directed Twitter links do not necessarily represent friendship ties but sometimes merely interest in the information produced by the followee. This leads to a denser link structure than in traditional social networks. As such, a follower network provides a middle ground between traditional broadcast media distribution (some nodes represent media outlets with millions of followers) and a more personal information exchange. Recent research has demonstrated that many follower links are actually reciprocal [13], suggesting that a significant portion of the network actually corresponds to personal friendship ties. On the other hand, there are a number of extremely high fan-out nodes corresponding to media outlets, companies and prominent public figures. As a result, it is difficult to judge how individual influence propagates in the network by simply observing the network structure on its own. Instead this task requires understanding of the behavior of nodes.

With regards to population-level dynamic behavior on a network, the spread of information on a network has been primarily expolored using models adopted from epidemiology [8,9], and have been applied to describe propagation rates of memes (i.e., Twitter hashtags) in social media [22]. We adopt these methods of analysis to evaluate the population-level topic behavior on influence networks, and assume a simple contagion model as the underlying propagation process in our data sets.

3 Genotype Model

Here we define our genotype model capturing the topic-specific behavior of a single user (node) within a social media network. Our main premise is that, based on observed network behavior, we can derive a consistent signature of a user. Hence, the genotype model is an individual user model, by definition, in the sense that it represents the behavioral traits of a social network user. For our analysis, the genotype

captures adoption and reposting of new information, activity levels, and latency of reaction to new information sent by influential neighbors. Other behavioral traits can be incorporated as well. The genotype is topic-specific as we summarize the behavioral traits with respect to a set of predefined topics.

A social media network $N(U, E)$ is a set of users (nodes) U and a set of follow links E. A directed follow link $e = (u, v), e \in E$ connects a source user u (*followee*) to a destination user v (*follower*). The network structure determines how users get exposed to information posted by their followees. The static network does not necessarily capture influence as users do not react to all information to which they are exposed. To account for the latter, we model the behavior of individual users taking into account their context in the follower network.

In its most general form, a user's genotype G_u is an entity embedded in a multi-dimensional feature space that summarizes the *observable behavior* of user u with respect to different topics. It is up to the practitioner to define the different dimensions of the topic feature space and the relevant aspects of observable behavior in the network locality of a node. Each genotype value can be viewed as an allele that the user introduces to the process of message propagation through a network.

In our study, we focus on hashtag usage within Twitter, since hashtags are simple user-generated tokens that annotate tweets generated by either a social group or designating a specific social phenomenon, and are often "learned" from others on the social network [1]. In this context, a hashtag serves as a genetic parcel of cultural information, just like alleles of a gene within a biological context. Hashtags can be associated with topics such that an individual's response to a collection of hashtags within a topic indicates a user's propensity to respond to other hashtags within that same topic.

We consider a finite set of hashtags $H = \{h\}$, each associated with a topic $T_i \in T$. To obtain the genotype, we analyze the social media message (tweet) stream produced by a user u, with respect to H. Let us define $m(\cdot)$ to be a function that maps each occurrence of (u, h) to a real values $m : \{(u, h)\} \mapsto \mathbf{R}$. The set of hastags associated with topuc T_i and adopted by user u are denoted as $H_{(u,T_i)} := \{h\}_{T_i} \cap \{h\}_u$, where $\{h\}_{T_i}$ is the set of hastags in topic T_i and $\{h\}_u$ is the set of hastags adopted by user u. The i^{th} element of the user genotype G_u is the set of $\{m(u, h) \mid h \in H_{(u,T_i)}\}$ values. We remark that this set of values may also be reduced to their average value or some approximated distribution function if one wishes to have a coarser representation of the data.

To construct each user's topic-genotype from empirical data, we consider a variety of metrics $m(\cdot)$ for (u, h) pairs, listed in Table 1. These metrics serve the purpose of quantifying a user's response to a hashtag by defining the data values that are used to estimate the topic distributions. While TIME and N-USES are intuitively obvious metric choices, LAT and LOG-LAT are novel to this manuscript. N-PAR and F-PAR have been previously studied in a different context [6], and are included here for comparison.

While we define the user genotypes based on adoption of hashtags in Twitter similar models can be built in other networks as well. The follower network structure in Twitter forms a directed graph and hence the definition can be easily generalized to undirected networks such as those of systems like Facebook and Google+. Instead of

Metric	Function definition	Notes
Time	TIME$(u,h) = min_{(u,h)}(t(u,h)) - min_{v \in V_u}(t(v,h))$, where $t(u,h)$ is the time (u,h) occurs and V_u is the set of followees of u.	The absolute amount of time between a users first exposure to the given hashtag and his first use of that same hashtag.
Number of Uses	N-USES$(u,h) = \lvert\{(u,h)\}\rvert$, where $\lvert \cdot \rvert$ is the cardinality function.	The total number of occurrences of the (u,h) pair.
Number of Parents	N-PAR$(u,h) = \lvert\{v \in V_u \mid t(v,h) < t(u,h)\}\rvert$	The number of followees to adopt before the given user.
Fraction of Parents	F-PAR = N-PAR$(u,h)/\lvert V_u \rvert$.	The fraction of a user's followees who have adopted the hashtag prior to the user.
Latency	LAT$(u,h) = \left(\lvert\{h_j \in H_{T_i} \mid H_{T_i} \ni h \text{ , and } t(u,h_j) < t(u,h)\}\rvert\right)^{-1}$.	The inverse of the number of same-topic hashtags posted to the user's time-line between his first exposure to the hashtag and his first use of the hashtag.
Log-latency	LOG-LAT$(u,h) = log\left(\text{LAT}(u,h)/Avg(\text{LAT}(w,h) \text{ s.t. } w \in U)\right)$	The logarithm of each latency value after each latency value has been divided by the mean latency value for that hashtag.

Table 1: Behavior-based metrics that are components of the topic-specific user genotype.

	SNAP (users=42M,tweets=467M)			CRAWL (users=9K,tweets=14.5M)		
Topic	Hashtags	Users	Uses/HT	Hashtags	Users	Uses/HT
Business	27	20k	1,155	19	1,493	88
Celebrities	32	26k	1,009	-	-	-
Politics	485	349k	2,020	121	5,480	49
Sci/Tech	33	415k	6,889	63	4,982	100
Sports	98	76k	3,274	24	320	14

Table 2: Statistics of the SNAP and CRAWL data sets.

hashtags one can focus on other aspects of behavior such as adoption of new phrases, hyper-links or other tokens that carry topical information.

4 Datasets

We chose Twitter to analyze user behavior via our genotype model since Twitter has millions of active users and messages have a known source, audience, time stamp and content. Similar analysis can be performed in other social media networks with a known follower structure and knowledge of the shared content (memes, URLs or buzz words) in time.

4.1 Twitter follower structure and messages

We use two datasets from Twitter: a large dataset SNAP [11] including a 20% sample of all tweets over a six-month period and the complete follower structure [5]; and a smaller CRAWL dataset containing all messages of included users that we collected using Twitter's public API in 2012, where we started from initial seed nodes (members of the authors' labs) and crawled the follower structure and related posts. SNAP includes a network-wide view for a 6 month period, while CRAWL provides longitudinal completeness for a smaller subnetwork of users. Statistics of the two datasets are summarized in Table 2.

The SNAP dataset contains 467 million posts from June to December 2009. The follower structure is based on the complete follower crawl of Kwak et al. [5] includ-

ing over 42 million Twitter users. CRAWL contains 14.5 million Twitter posts from March 2006 to May 2012. The CRAWL follower structure was obtained at the time of crawling the tweets and includes $9,468$ users and 2.5 million follower links (the number of links includes followees for whom we do not have tweets). Due to its size, CRAWL has a sparser hashtag representation (e.g. no hashtags from our curated list of Celebrity-related hashtags). We reproduce our experiments in both datasets in order to evaluate the effect of sub-sampling the messages in SNAP. The behavior and utility of our genotype model is persistent for both datasets.

The data values for each genotype metric are likely to be affected by the fact that 80% of the SNAP users' messages were not recorded. In addition, not all hashtags we encounter can be attributed to a topic. Nonetheless, all metrics in this study are affected equally, and evaluated relative to each other. Obtaining complete snapshots of network structure at any given point in time in these experiments is untenable. Thus, we acknowledge this limitation and cast our results in the context of only what is known about the network structures and posts within the respective datasets.

4.2 Grouping hashtags into topics

While hashtags present a concise vocabulary to annotate content, they are free-text user-defined entities. Hence, we need to group them into topics in order to summarize user behavior at the topic level. In this work we assume each hashtag belongs to exactly one topic, while in a more general framework disseminated hashtags (URLs, memes, etc.) can be "softly" assigned to more than one topic. We work with five general topics as dimensions for our user genotypes: Sports, Politics, Celebrities, Business and Science/Technology. We obtain a set of 100 hashtag annotations from a recent work by Romero and colleagues [6], further augmented by a set of curated business-related hashtags [23]. We combine this initial set of annotated hashtags with a larger set based on text classification.

To increase the number of considered hastags, we adopt a systematic approach for annotating hashtags based on URLs within the tweets. To associate tweets with topics, we treat user-generated hashtags as tokens that carry topical identity, similar to previous studies [6]. Users include hashtags to annotate (topically) their tweets and to participate in a specific community discussion [11]. Adopting the appropriate hashtag for a message ensures better chances of surfacing the content in search as well as attracting the attention of interested followers.

We pair non-annotated hashtags with web URLs, based on co-occurrence within posts. We extract relevant text content from each URL destination (most commonly news articles from foxnews.com, cnn.com, bbc.co.uk) and build a corpus of texts related to each hashtag. We then classify the URL texts in one of our 5 topics using the MALLET [24] text classification framework trained for our topics of interest. In order to train the MALLET [24] topic classifier we use annotated text from two widely used topic-annotated text collections: the 20 newsgroups dataset [25] and the News Space [26]. Additional ground-truth text collections can be used for wider topic coverage and to improve the accuracy.

As a result, we get a frequency distribution of topic classification for frequent (associated with at least 5 texts) hashtags. The topic annotation of the hashtag is the topic of highest frequency. The number of hashtags and their usage statistics in our final topic-annotated set are presented in Table 2 (columns Users and Uses/HT). The Celebrities hashtags do not occur frequently enough in the CRAWL dataset and hence we exclude them from our analysis.

4.2.1 Discussion

Alternative information retrieval and natural language processing approaches for annotating tweets into topics can also be adopted within our framework. Hashtags, as a means of annotation and defining a universal vocabulary, are also common in systems for other types of content such as music, photos and video. Examples include the photo sharing social site Flickr, the video sharing site Youtube, and music streaming sites such as Last.fm and Pandora. We believe that our hashtag-based genotype framework might extend to modeling and analysis of user behavior when interacting and disseminating photos and multimedia as well.

We adopt a model in which every information item (hashtag) is associated with exactly one topic. This particular way to instantiate our genotype model is the first attempt to demonstrate the preserved behavior within a topic. One can naturally extend this to a richer analysis in which we have "soft" association of content items and topics. One promising direction is to learn such association using latent topic models such as the ones introduced by Blei and colleagues [27] in lieu of hard topic classification. Our proposed applications (topic prediction, latency minimization, and adoption prediction) can then be extended naturally using the probabilistic association weights of hashtags for different topics.

5 Genotype model validation in Twitter

To justify the genotype model as a meaningful representation of social network users, we demonstrate that it is capable of capturing stable individual user behavior for a given topic. We seek to evaluate the stability of configuration of multiple users' genotype values within a topic, and use a classification task and the obtained (training/testing) accuracy as a measure of consistency for our genotype model. Within this context, we compare different genotype dimensions and evaluate the level to which each of them captures characteristic invariant properties of a social media user.

5.1 Topic consistency for individual users

Our hypothesis is that individual users exhibit consistent behavior of adopting and using hashtags (stable genotype) within a known topic. If we are able to capture such invariant user characteristics in our genotype metrics, then we can turn to employing the genotypes for applications. We compute genotype values according to our collection of hashtags with known topics by training a per-user Linear Discriminant

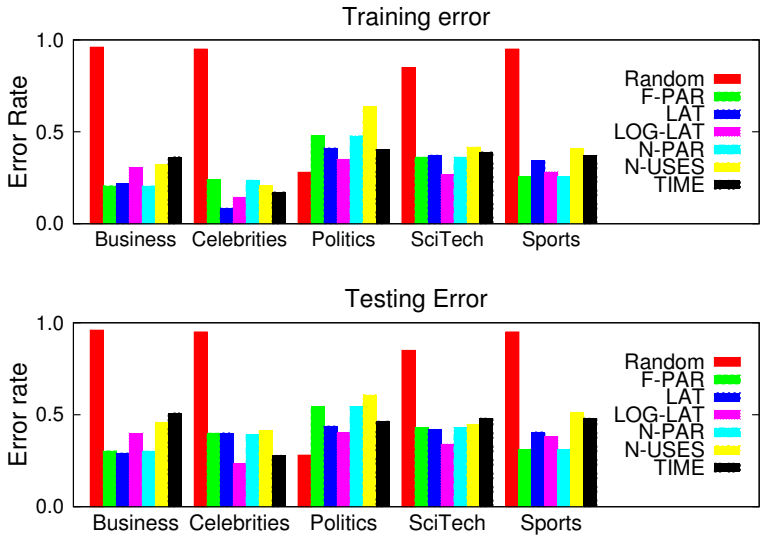


Fig. 1: Training and testing accuracy of hashtag classification in a leave-one-out Linear Discriminant classification.

(LD) topic classifier to learn the separation among topics. Consider, for example, the LOG-LAT genotype metric: for a user u, we have a set of observed LOG-LAT values (based on multiple hashtags) that are associated with the corresponding topics. If the user u is consistent in her reaction to all topics, then the LOG-LAT values per topic will allow the construction of a classifier with low training and testing error. It is also noted that each hashtag does end up having a topic distribution, but for the scope of this study, a sufficient hashtag classification should at least agree in the topic of greatest probability/likelihood, which is what is presented here.

The consistency of user responses is evaluated using a leave-one-hashtag-out validation. Given the full set of (u, h) response values, we withhold all pairs including a validation hashtag h and employ the rest of the pairs involving hashtags of known topic to estimate the individual user's topic genotype. We repeat this for all genotype metrics. The training and testing error rate for this experiment are presented in Fig. 1, and their similar error rates demonstrate how consistent users are at classifying hashtags into topics. In both cases, our genotype metrics enable significantly lower error rates than a Random model (i.e. random prediction based on number of hashtags within a topic), demonstrating that, in general, genotype metrics capture consistent topic-wise behavior. One exception is the Politics topic as it has comparatively many more hashtags than other topics, skewing the random topic distribution resulting in slightly lower error. Across genotype metrics, we observe that normalized latency of adoption (LOG-LAT) is more consistent per user than alternatives.

	Bus.	Celeb.	Poli.	Sci./Tech.	Sport	E[x]
Random Error	0.96	0.95	0.28	0.85	0.95	0.45
F-PAR	0.50	0.88	0.61	0.15	0.09	0.41
LAT	0.09	0.46	0.18	0.19	0.25	0.21
LOG-LAT	0.05	0.13	0.19	0.12	0.03	0.13
N-PAR	0.09	0.50	0.88	0.09	0.03	0.40
N-USES	0.45	0.42	0.90	0.22	0.56	0.54
TIME	1.0	1.0	0.01	0.92	0.88	0.61

Table 3: Error rates of the NB consensus topic classification. $E[x]$ is the expected error across topics.

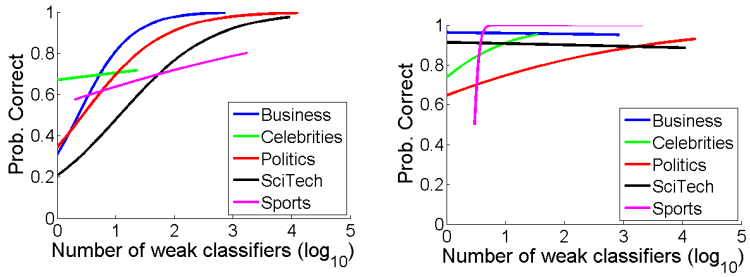
5.2 Topic consistency within the network

While individual users may exhibit some inconsistencies in how they behave with respect to hashtags within a topic, an ensemble of users' genotypes remains more consistent overall. To demonstrate this effect, we extend our classification-based evaluation to the network level. We implement a network-wide ensemble-based Naive Bayes (NB) classifier that combines output of individual user classifiers to achieve network-wide consensus on the topic classification of each validation hashtag.

To implement a Naive Bayes consensus classifier on the output of each user's local LD classifier, posterior topic distributions are required for each topic of each user's genotype. We assume normality for these distributions within each topic, where the mean values are centered about the correctly classified training hashtags and the variance is computed from all training hashtags for that topic. The topic prior distributions are estimated from the relative proportion of hashtags in each topic, and the hashtag's ultimate topic classification is determined by the maximum posterior likelihood over the network (all user-wise LD classification outputs).

Table 3 summarizes the testing error rate of our NB scheme for classifying hashtags into topics in a leave-one-hashtag-out validation. The consensus error rate decreases compared to local classifiers (Fig. 1), demonstrating that the genotypes, as a complex, are more stable and consistent than individual users. The lowest error rate of 0.13 is achieved when using the LOG-LAT metric. The TIME metric happened to be the least accurate metric of them all, because individual user response time values (TIME) showed the least discernable clustering behavior. The accuracy of the TIME metric performed most similar to the null (Random) model when compared the other metrics on a topic-by-topic basis, but TIME happened to be more biased towards political hashtags because they occurred most frequently in the dataset.

The latency genotype metrics that are most invariant (LAT and LOG-LAT) implicitly normalize their time scales of response with respect to the user's own frequency of activity, which is a feature not captured by the absolute TIME metric, or any of the other metrics. Furthermore, both of these metrics incorporate the network structure, measuring the message offset since the earliest exposure to the hashtag via a followee. LOG-LAT has a slight advantage over LAT because it suppresses the background noise of each hashtag measurement. However, LOG-LAT has the disadvantage of being dependent on a network-wide latency measurement for the same hashtag, which might be harder to obtain in practice. In this sense, LAT is a



(a) LAT Net Classifier (b) LOG-LAT Net Classifier

Fig. 2: Accuracy of the network classification as a function of the number of local classifiers (SNAP). A logistic function is fit to each topic's accuracy.

more practical genotype dimension when summarizing individual user behavior in real time.

While the system of all user genotypes exhibits significant consistency (high classification accuracy), it is useful to know how many user genotypes are needed to obtain a good classification (i.e. detect a network-wide topic-specific spread). We observe an increasing classification accuracy with the number of users included in the NB scheme. Figures 2a and 2b show the dependence of accuracy on number of local LD classifiers included per topic. All curves increase sharply, indicating that variability within individuals is easily overcome by considering a small subset of users within the network. In fact, the Business and Sci./Tech. accuracies in Figure 2 are most accurate for the smallest subset of users (i.e., fewest number of local classifiers), and then decrease slightly as less reliable individuals are included in the network classifier. Overall, the accuracy of the LOG-LAT network classifier tends to increase *faster* to its optimal level with increasing number of local classifiers, since the LOG-LAT metric features a network wide normalization and thus contains global information.

With increasing number of available individual genotypes, the Business topic requires consistently fewer local classifiers than the Celebrities. One explanation of this might be a higher heterogeneity of sub-topics within Celebrties and hence lower topic-wide response consistency. For example, many businesses and brand names are designed to be topically distinct, while celebrities may be perceived as sports stars, politicians, or company executives. For topics like the latter, more individual genotypes are needed to arrive at a correct hashtag classification.

It is important to note that we use classification only as a way to evaluate if the topic specific-behavior captured by our genotype metrics is invariant for users. While the genotypes might be adopted for actual novel information classification into topics, an improved classifier for such applications may benefit from combining the genotypes with textual features of tweets.

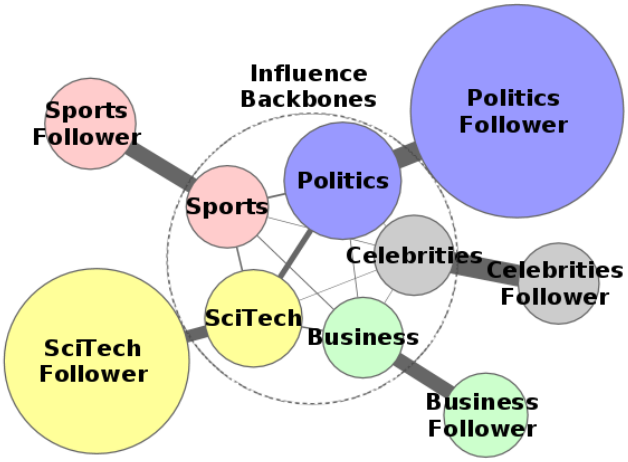


Fig. 3: Overlap among topic influence and corresponding follower subnetworks (in SNAP). Each network is represented as a node, with every topic represented by an influence (encircled in the middle) and a follower network. Node sizes are proportional to the size of the network (ranging from $120k$ for Celebrities to $42m$ for Politics Follower). Edge width is proportional to the Jaccard similarity of the networks (ranging from 10^{-3} inter-topic edges to 10^{-1} between corresponding influence-follower networks).

6 Topic-specific influence backbones

As we demonstrate in the previous section, user behavior remains consistent within a topic. A natural question inspired by this observation is whether topics propagate within similar regions of the shared medium that is the follower network structure. By observing the behavior of agents (adoption, reposting, etc.) one can reveal the underlying backbones along which topic-specific information is disseminated. In this section, we study the propagation of hashtags within Twitter to identify *topical influence backbones* — sub-networks that correspond to the dynamic user behavior. We superimpose the latter over the static follower structure and perform a thorough comparative analysis to understand their differences in terms of structure and population-level user behavior. The topical backbones in combination with the individual user genotypes will then enable various applications as we show in the subsequent section.

6.1 Influence backbone definition and structure

An *influence edge* $e_i(u, v)$ connects a followee u who has adopted at least one hashtag h within a topic T_i before the corresponding follower v. Hence, the influence network $N_i(U, E_i)$ for topic T_i is a subnetwork of the follower network $N(U, E)$ (including the same set of nodes U and a subset of the follower edges $E_i \in E$). We weight the

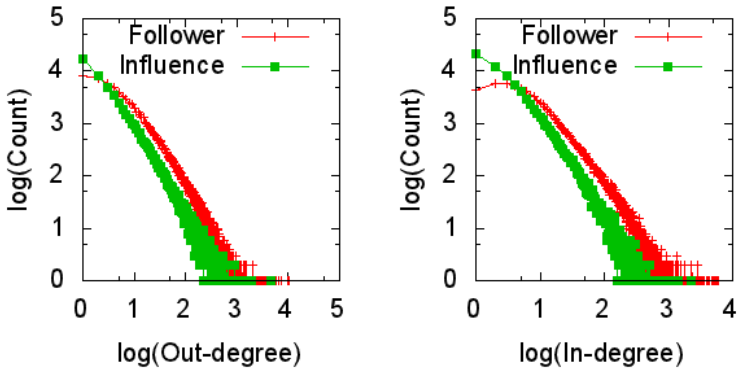


Fig. 4: Out- and In-Degree distributions for the Follower and Influence networks for *Sports* (SNAP).

edges of the influence network by the number of hashtags adopted by the followee after the corresponding follower, and within the same topic.

First, we seek to understand the differences between the influence backbones and the static follower network. Figure 3 presents the overlap among influence backbones and their corresponding follower network. For this comparison, we augment an influence network with all follower edges among the same nodes to obtain the corresponding follower network. In the figure, each network is represented by a node whose size is proportional to the network size (in edges). Connection width is proportional to the *Jaccard Similarity (JS)* (measured as the relative overlap $|E_i \bigcap E_j|/|E_i \bigcup E_j|$) of the edge sets of the networks. The Jaccard similarity for influence and follower networks varies between 0.16 for Sports to 0.3 for Celebrities. The influence networks across topics do not have high overlap (*JS* values not exceeding 0.01), with the exception of Sci/Tech and Politics with $JS = 0.07$. This may be explained partially by the fact that these are the largest influence networks (5 and 11 million edges respectively). Another reason could be that there are some "expert" nodes who are influential and active in both topics.

The degree distributions of influence and follower networks within a topic maintain a similar shape. Figure 4 shows the in- and out-degree distributions for the *Sports* networks (in SNAP). The most dramatic change in the distributions is for small degrees with almost one magnitude increase of the nodes of in-degree 1. Users who retain only a few influencers tend to have a variable number of followees, hence the in-degree distribution decreases for the whole range of degrees.

Beyond network sizes and overlap, we also quantify the structural differences of the influence backbone in terms of connected components. A *strongly connected component (SCC)* is a set of nodes with directed paths among every pair, while in a *weakly connected component (WCC)* connectivity via edges regardless of their direction is sufficient. Figure 5 compares the sizes of the largest SCC and WCC in the topic-specific networks as a fraction of the whole network size. When ignoring the direction (i.e. considering WCC), both the influence and follower structures have a single large component amounting to about 99% of the network. The communities

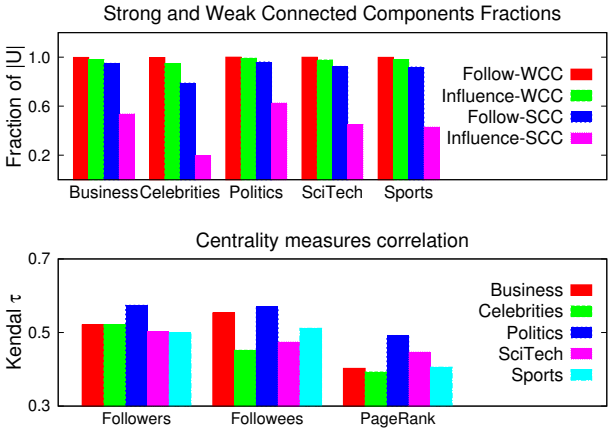


Fig. 5: Largest weakly and strongly connected component (WCC and SCC) sizes as a fraction of the network size (top); and Kendall τ rank correlation of node importance measures for the influence and follower networks (bottom) (SNAP).

that are active within a topic are connected, showing a network effect in the spread of hashtags, as opposed to multiple disjoint groups which would suggest a more network-agnostic adoption. When, however, one takes direction into consideration (SCC bars in Fig. 5), the size of the SCC reduces drastically in the influence backbones. Less directed cycles remain in the influence backbone, resulting in a structure that is close to a directed acyclic graph with designated root sources (first adopters), middlemen (transmitters) and leaf consumers. The reduction in the size of the SCC is most drastic in the Celebrities topic, indicative of a more explicit traditional media structure: sources (celebrity outlets or profiles) with a large audience of followers and lacking feedback or cyclic influence.

How does a user's importance change when comparing influence to following? In Figure 5 (bottom) we show the correlation of node ranking based on number of followers, followees and PageRank [28] in the influence and follower networks. The correlation of each pair of rankings is computed according to the *Kendall* τ rank correlation measure. The correlation is below 0.5 for all measures and topics. Global network importance (PageRank) is the most distorted when retaining only influence edges (0.4 versus 0.5 on average), while locally nodes with many followers (or followees) tend to retain proportional degrees in the influence network.

While the follower structure features a lot of reciprocal (bi-directional) links (above 50% on average), these reciprocal links disappear almost completely in the influence backbone (retaining 4% on average), as shown in Fig. 6. This effect is most prominent in the Celebrities topic where reciprocal links drop from 36% to less than 1% in the influence network. Reciprocal links are related to friendship ties, i.e. nodes who are possibly friends declare interest in each other's posting by a bi-directional link. When it comes to influence, however, the ties tend to be uni-directional with only one of the nodes affecting the other.

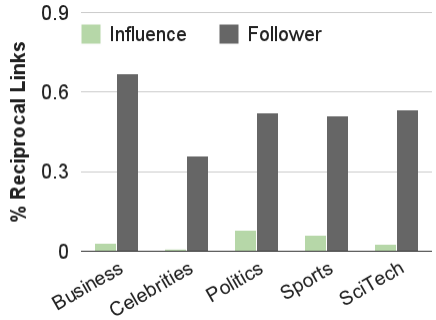
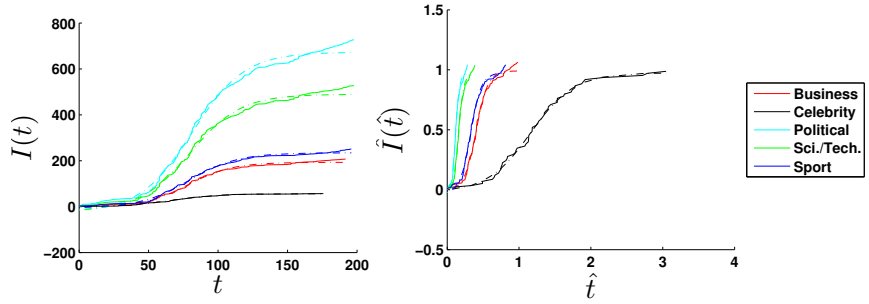


Fig. 6: Comparison of the percentage of reciprocal (bi-directional) links in the influence and follower networks



(a) (b)

Fig. 7: Example of typical regression result, from data of the Political hashtag *#beck*, referring to the political commentator Glenn Beck. (a) The measured data (solid lines) and the approximated regression function (dashed lines) in the unnormalized coordinates, and (b) the same data in the normalized coordinates. The plotted curves are colored according to the topic backbone that the *#beck* hashtag was detected on.

Our comparative analysis of the influence and follower structure demonstrates that the influence backbone is quantitatively different from the overall follower network. The explanation for this lies in the fact that the influence backbone is based on the dynamic behavior of users (information dissemination on specific topics), while the follower structure represents the static topic-agnostic media channels among users. Not all followees tend to exert the same amount of influence over their audiences in the actual information dissemination process, giving rise to distinct topic-specific influence backbones. We obtain similar behavior in the smaller Twitter data set CRAWL (omitted due to space limitation).

6.2 Population behavior on topic backbones

Thus far, the topical influence backbone networks are comprised of the individuals responsive within a given topic. Additionally, the results at the individual scale, as described in Section 5, demonstrate aggregate consistency among users for how they behave towards hashtags of similar topic. Since many users are members of more than one backbone, yet may be more responsive towards one topic than another, an ensuing question is whether dynamics on the topic backbones are consistent with individual behavior. *Does the Business backbone, for example, propagate business hashtags faster than, say, the Sports backbone?* In general, we find this hypothesis to be true, assuming that the underlying hashtag propagation process follows a simple epidemic-inspired compartmental population model.

Compartmental population models are often implemented to study average behavior of a disease or meme within a population [8,9,22]. In the simplest case where we have only two classes of individuals, susceptible (S) and informed (I), a susceptible individual can become informed of a meme, and once informed will remain informed. Such coarse two-state models for simple contagions (i.e., cascades) describe average rates of adoption from one class of individuals to the next. For static populations, where $S + I = N$ for some fixed population of size N, the dynamics of a typical S-I process are defined by Newman et al [9] as:

$$\frac{dI}{dt} = \beta I(N - I), \tag{1}$$

which has the solution

$$I(t) = \frac{NI(0)e^{\beta t}}{N + I(0)\left(e^{\beta t} - 1\right)}, \tag{2}$$

where β is the transmission rate and $I(t)$ is the size of the infected population at time t.

One can quantify and compare the contagiousness of a hashtag on different networks by comparing its respective β values. An example set of realizations is depicted in Figures 7a and 7b. It is important to note the sigmoidal shape of the adoption curves and their least-squares approximations. This sigmoidal shape is characteristic of processes governed by Eq. (2).

For this particular study, we track a hashtag of known topic on the Twitter network in order to observe whether or not the hashtag is most *viral* on its own topic backbone. We begin by considering only hashtags that have been tweeted by users who are members of more than one topic backbone within the SNAP dataset. A distinct realization of Eq. (1) for a hashtag is defined by the total population of individuals who have tweeted that hashtag with respect to time.

When comparing the model defined by Eq. (2) to temporal hashtag data, one needs to account for the fact that the hashtag may have existed on the network prior to the time of initiating data acquisition. Hence, the first observed use of a hashtag in our data is possibly not the actual first use of that hashtag. To account for this uncertainty of initial hashtag usage time, we shift the initial tweet of each hashtag to the origin by an amount of time τ, such that $I(0) = 1$ in all cases, and add a variable I_{t-} to account for the existence of an informed population before the first

hashtag detection. Therefore, Eq. (2) becomes a regression problem with four degrees of freedom: N, β, τ, and I_{t-}. The least-squares objective function is defined as

$$\text{minimize} \sum_i |y(t_i) - I(t_i)|^2 \tag{3}$$

for all i data points of the given hashtag. Here, $y(t_i)$ are the observed data points, and $I(t)$ is given by

$$I(t) = \frac{Ne^{\beta(t-\tau)}}{N + \left(e^{\beta(t-\tau)} - 1\right)} - I_{t-}. \tag{4}$$

Since Eq. (4) requires a count of only the total population for $I(t)$ rather than the specific backbone network topology, the backbones are used to identify the subset of topic users whose collective hashtag adoption makes each $I(t)$ signal. The N, β, τ, and I_{t-} parameters are deduced from a non-linear least-squares regression of Eq. (4) on the set of $(t, I(t))$ points for each hashtag realization on a backbone network.

For each hashtag h that is tweeted on more than one topic backbone B, there exists a transmission rate parameter $\beta(h)$ and effective population size $N(h)$ for each of those backbones. In order to compare the $\beta(h)$ parameters for backbones of different effective population sizes, we must first normalize each $I(t)$ signal with respect to its best fit $N(h)$. By factoring N out of the right-hand side of Eq. (1) and dividing both sides of Eq. (1) by N, one obtains

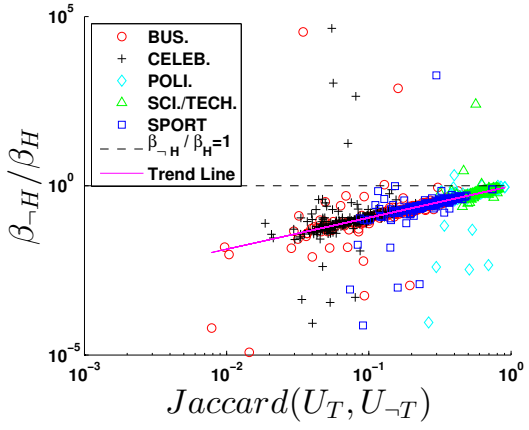
$$\frac{d\hat{I}}{d\hat{t}} = \hat{\beta}\hat{I}(1 - \hat{I}), \tag{5}$$

where $\hat{\beta} = \beta N$ and $\hat{I} = I/N$. It is also noted that substituting $\beta = \hat{\beta}/N$ into Eq. (2) leads to the normalized time scale $\hat{t} = t/N$. In this normalized setting, one interprets $\hat{\beta}$ as the number of interactions per unit of time (i.e., tweets among individuals that contain the hashtag of interest).

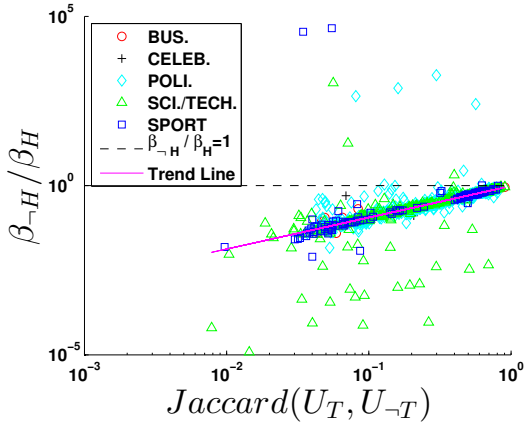
There are many hashtag users who are present on more than one topic backbone such that when one of these individuals uses a hashtag, that hashtag is observed to be simultaneously propagating on each topic backbone to which the user belongs. For example, suppose a Business related hashtag is used by an individual who is a member of the Business, Politics, and Sports topic backbones. The true topic (T) of this particular hashtag is Business, and a not true topic $(\neg T)$ is either Politics or Sports. In this case, there will be two $(T, \neg T)$ pairs: (Business,Politics) and (Business,Sports).

We denote the transmission rate of the hashtag on its actual topic backbone $\hat{\beta}_T(h)$ and the hashtag transmission rate on an off-topic backbone as $\hat{\beta}_{\neg T}(h)$. For each hashtag, we also denote the Jaccard similarity between the subset of those hashtag users on the backbones of a $(T, \neg T)$ pair as Jaccard$(U_T(h), U_{\neg T}(h))$, where $U_T(h) := \{u \in B_T \mid \forall u \in (U, h)\}$ and $U_{\neg T}(h) := \{u \in B_{\neg T} \mid \forall u \in (U, h)\}$. Recall that B represents the topic backbone, and should not be confused with β, which represents the transmission rate of Eq. (1).

Figures 8a and 8b show the data comparing $\hat{\beta}_T(h)$ relative to each $\hat{\beta}_{\neg T}(h)$ in the vertical dimension, and the Jaccard similarity of the respective users of h in the corresponding T and $\neg T$ backbones, in the horizontal dimension. Overall, we see



(a)



(b)

Fig. 8: Relative transmission rate with respect to Jaccard similarity between two backbones on which a hashtag propagates in the SNAP dataset. The same data points are shown in both (a) and (b), but with different marking schemes, and each point in either plot represents a $(T, \neg T)$ pair. Color is added to improve marker differentiation. (a) Colors indicate the topic backbone on which a given hashtag h is propagating (i.e., colored by the $\neg T$ topic). (b) Colors indicate the true topic to which the given hashtag h belongs (i.e., colored by the T topic).

that, on average, each hashtag propagates fastest on its own topic network since an overwhelming majority of the data points lie below the $\hat{\beta}_{\neg T}(h)/\hat{\beta}_T(h) = 1$ line.

Figure 8a demonstrates that the relative rates of propagation tend to increase as the topic backbones increasingly overlap. This is particularly evident for the Business, Celebrity, and Sports topic backbones. The collection of Sci./Tech points below the trend line of Figure 8b indicates that these hashtags have transmission rates on

off-topic backbones $\hat{\beta}_{\neg T}(h)$ that are much less than their true topic backbone $\hat{\beta}_T(h)$. The corresponding points in Figure 8b indicate which off-topic backbone yields the transmission rate $\hat{\beta}_{\neg T}(h)$.

Outliers in Figs. 8a and 8b are an artifact of the SI-model not being an appropriate underlying model for their data, but are included in the results because either the T or $\neg T$ backbones for the associated hashtag proved to have SI-type behavior. The outliers, however, have little effect on the trend line shown in Figures 8a and 8b, since the trend line has an average point-wise residual of 0.15 on the log-log scale shown.

7 Applications of genotypes and backbones

In this section, we employ the user genotypes and the topic-specific backbones for two important applications: (i) prediction of hashtag adopters and influencers and (ii) latency minimization of topical information spread. In both applications knowledge of individual genotypes and influence backbones enables superior performance compared to the static network structure on its own.

7.1 Topic-specific influence prediction

We employ the influence structure and the user genotypes to predict likely influencers/adopters for a hashtag. We aim to answer the following question: *Which followees are likely to influence a given user to adopt a hashtag of a certain topic and analogously which followers are likely to adopt a hashtag?* This question is of paramount importance from both research and practical perspectives. On one hand, uncovering the provider-seeker influence will further our understanding of the global information network dynamics. On the other hand, the question has practical implications for social media users offering guidelines on following high-utility sources or keeping the follower audience engaged.

In this experiment, we consider *(u,h)* pairs, for users who have at least 10 followees and have used the hashtag at least once. The goal is to predict the subset of all followees who have used the hashtag prior to the user in question and similarly all adopting followers who are likely to use the hashtag later. We construct three structural predictors utilizing the follower structure that rank influencers/adopters by *Followees*, *Followers* and *Reciprocal* links. Ties are broken in a random manner.

The genotype-based predictors utilize genotypes and influence edges to rank influencers and adopters. This group of predictors includes ranking by (i) topic-specific activity in terms of number of usages of hashtags within the same topic (dimension N-USES of the genotype) as the target tag (*Topic Act*); (ii) general tweeting activity involving all hashtags (*Act*); and (iii) a predictor that combines the activity and the influence backbone (*RW+Act*). *RW+Act* performs random walks in the influence backbone for the same topic and considers the probability of visit of the followees/followers of the target as a weight of the candidate. Ties in the probabilities (certain to arise when candidates are isolated after removing the target hashtag

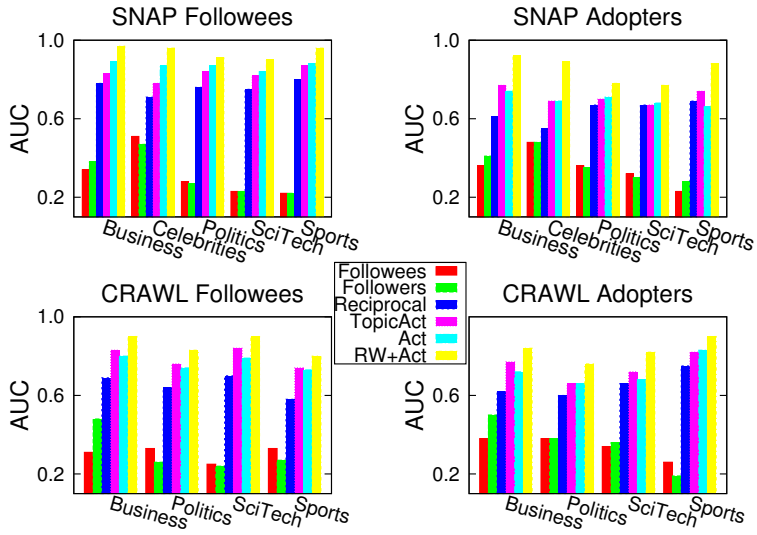


Fig. 9: Influential followee and adopter prediction accuracy. We consider several predictors of a user's influencers by a hashtag in a known topic. Genotype-based predictors (*Act*, *Topic Act* and *RW+Act*) perform better than follower structure-only counterparts (Follower, Followee and Reciprocal).

or lacking influence links altogether) are broken by using the topic-based activity. When a tie is observed in topic-based activity as well, the overall activity is used for ranking. One can view the *RW+Act* predictor as combining *Act*, *Topic Act* and a random walk importance measure in the influence network. None of the predictors has information about the spread of the specific hashtag.

A prediction instance is defined by a user u and an adopted hashtag h. Only a subset $I(u,h)$ of all structural followees/followers of the user are true influencers/adopters (positives for the prediction task). Our goal is to predict the subset of true influence neighbors using their features and local influence structure (excluding information about the same hashtag h). A good predictor ranks the true neighbors first. In order to overcome the effect of sparsity in the data, we consider prediction of instances for which at least one candidate followee is not isolated in the influence network after removing the links associated with the target hashtag. We measure true positive and false positive rates for increasing value of k (the maximal rank of predicted influencers/adopters) and compute the average area under the curve (AUC) as a measure of the predictor quality. We report this measure within each topic in Figure 9 for the SNAP (top) and CRAWL (bottom) datasets and for influence followees (left) and adopter (right) prediction. Overall, in both datasets, the genotype-based predictors outperform the structure-only counterparts. The existence of a reciprocal follow link is the best structure-only predictor implying the importance of bi-directional links which often may correspond to a friendship relationship [6]. Social friends have been found to re-share the same information with a very low latency in a recent large scale field experiment [29], which may also be related to reciprocal links perform-

ing closer to the genotype-based predictors as compared to number of followees or followers. The genotype-based predictors relying on topic specific activity, overall activity and the influence structure allow over 20% improvement with respect to the reciprocity predictor and above two-fold improvement compared to number of followees/followers predictors. Although node information alone (*Act* and *Topic Act*) provides a good accuracy, this effect is even stronger when combining them with the knowledge of the topic influence network in the composite *RW+Act* predictor. The *RW+Act* increases the rank of followees who have influenced the same user or other users within the same topic for different hashtags.

RW+Act's improvement is highest in the Business and Sports topics and lowest in Politics and SciTech for the SNAP dataset. This may be due to the fact that in Business and Sports there are highly topic-specialized authoritative users that followers pay attention to, while Politics and Sports constitute wider-spread topics that appeal to everyone, and hence followers tend to adopt them from their most active followees.

The predictor performance is similar in the CRAWL dataset (Fig. 9), showing the generalization of our models to different types of data. The smaller improvement in CRAWL (compared to SNAP) can be explained partially by sparser usage of analyzed hashtags or due to possibly evolving genotypes of users over longer time frames, a hypothesis we are planning to evaluate in future work.

7.2 Network latency minimization

Another important problem that can be addressed given knowledge of topic-specific user behavior is that of improving the speed of information dissemination. Fast information dissemination is critical for social-media-aided disaster relief, large social movement coordination (such as the Arab Spring of 2010), as well as time-critical health information distribution in developing regions. In such scenarios, genotypes and the influence structure among users are critical for improving the overall "latency" of the social media network. In this subsection, we demonstrate the utility of our individual user models for latency minimization.

Consider a directed path in a topic-specific influence backbone N, defined by a sequence of nodes $P = (u_1, u_2 ... u_k)$. The *path latency* $l(P)$ is defined as the sum of topic-specific latencies (*Time* measure of the genotype)

$$l(P) = \sum_{j=1...k-1} Time(u_j), u_j \in P$$

of all nodes except the destination. The *source-destination latency* (or just latency) $l(u_1, u_k) = \min_{P:u_1 \to u_k} l(P)$ is defined as the minimum path latency considering all directed paths between the target nodes. The concept of latency is similar to that of shortest path length, except that "length" is measured according to the responsiveness of traversed nodes (i.e., minimal time until u_k's adoption of a hashtag introduced by u_1). The *average network latency* is defined as the mean of all node pair latencies. Given a directed network $N(V, E)$ and latency for every node, we define the problem of k *Latency Minimization (k-LatMin)* as finding the k best target nodes, whose latency reduction leads to the largest average network latency decrease. We assume that

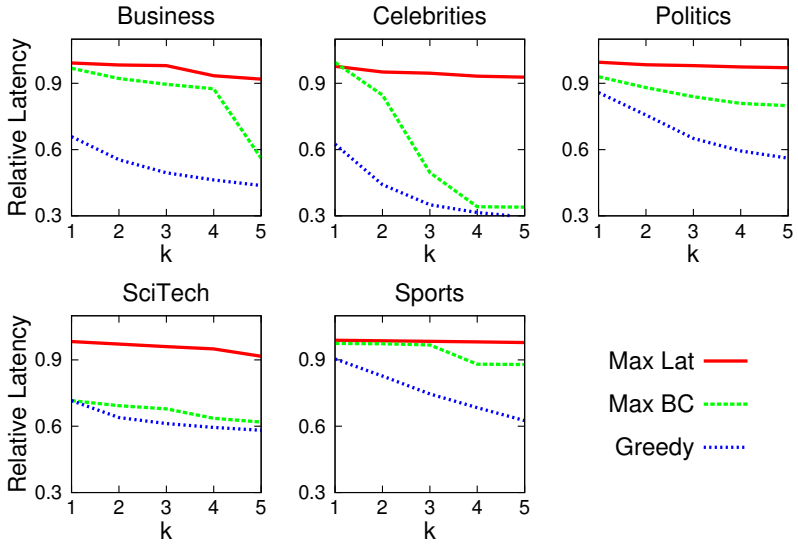


Fig. 10: Comparison of three heuristics for Latency Minimization in the SNAP dataset. The traces show the relative (w.r.t. the original) average network latency as a function of the number of targeted nodes k.

specific nodes could be targeted to reduce their individual latency. In real application scenarios, node latency can be reduced by timely and relevant content recommendation to target nodes and/or financial incentives. For our analysis we optimistically assume that every node's latency could be reduced to 0, however, node-wise constraints can be incorporated according to known limitations of users.

One can show (via a reduction from the *Set cover* problem) that k-LatMin is NP-hard. We consider three heuristics: *Max Lat* targets nodes in descending order of their latency values; *Max BC* targets nodes in decreasing order of their structural node betweenness-centrality measure; and *Greedy* targets nodes based on their maximal decrease of average latency combining both structural (centrality) and genotype (latency) information.

Figure 10 shows the performance of the three heuristics in minimizing the average latency in subgraphs (of size 500 nodes) of the largest strongly connected components within the influence backbones of our SNAP dataset. Considering the node genotypes (Max Lat) or the influence backbone (Max BC) on their own is less effective than jointly employing both (Greedy) across all topics. The Greedy heuristic enables about 2-fold reduction of the overall network latency by targeting as few as 1% (5 out of 500 nodes) of the user population. It is interesting to note that in Sports and Celebrities, since there are central nodes of large degrees, the betweenness-centrality criterion performs almost as good as Greedy.

8 Discussion and future directions

The presented study is a first step towards characterizing topic-specific user behavior as genotypes, and our real-world analysis was focused on Twitter hashtags only. While this restriction leads to sparsity in the data and limited observations to construct the genotype dimensions and influence backbones, our goal was to demonstrate the utility of creating such a user behavioral model. More general information parcels can also be adopted in the future, including spread of URLs and topic sentiment. Establishing the minimum sufficient number of observations to obtain stable genotypes is another issue that needs to be addressed in the future.

Another future direction is investigating if genotypes change over time. While genotypes remain constant for a certain period, they might drift over longer periods of time, e.g. people developing new interests or changing political views. The slightly lower influence predictions in the longitudinal dataset CRAWL attests to such possibility, and we are planning to investigate the existence of drift in the genotype values as future research.

Important future challenges related to our latency minimization application include (i) a constant factor approximation algorithm, (ii) scaling up the *Greedy* approach (the current naive approach takes computation times on the order of hours for networks larger than 500 nodes); and (iii) considering cost-aware network manipulations by relating the utility of decreased latency to the cost of targeting nodes for real-world scenarios. In addition, one can also allow link addition manipulations similar to in the *average shortest path minimization (Min-SP)* problem [30].

9 Conclusion

We introduced the social media genotype—a genetically-inspired framework for modeling user participation in social media. Features captured by the user genotypes define the *actual topic-specific* user behavior in the network, while the traditionally analyzed follower network defines only *what is possible* in the information dissemination process. Within our genotype model, each network user becomes an individual with a unique and invariant behavioral signature within the topic-specific content dissemination. In addition, we demonstrated that users are embedded in topic-specific influence backbones that differ structurally from the follower network. Using a simple contagion model, these backbones were shown to propagate hashtags fastest when the backbone and hashtag belong to the same topic.

We instantiated our topic-based genotype and backbone framework within a large real-world network of Twitter and employed it for the tasks of *(i)* discovering topic-specific influencers and adopters, and *(ii)* minimizing the network-wide information dissemination latency. The genotype framework, when combined with the topic-specific influence backbones, enabled good influence predictive power, achieving improvement by more than 20% over using the follower structure alone. In the latency minimization application, we demonstrated that the knowledge of topic backbones and genotypes can enable 2-fold reduction of the overall network latency by reduc-

ing the latency of appropriately selected nodes that represent only 1% of the user population.

Acknowledgements. This work was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

1. O. Tsur and A. Rappoport, "What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities," in *WSDM*, 2012, pp. 643–652.
2. P. Bogdanov, M. Busch, J. Moehlis, A. K. Singh, and B. K. Szymanski, "The social media genome: Modeling individual topic-specific behavior in social media," in *ASONAM*, 2013.
3. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD*, 2003, pp. 137–146.
4. M. Kimura, K. Saito, R. Nakano, and H. Motoda, "Extracting influential nodes on a social network for information diffusion," *DMKD*, 2010.
5. H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *WWW*, 2010, pp. 591–600.
6. D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," in *WWW*, 2011, pp. 695–704.
7. N. Friedkin, *A Structural Theory of Social Influence*. Cambridge University Press, 2006, vol. 13.
8. H. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.
9. M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.
10. P. Dodds, K. Harris, I. Kloumann, C. Bliss, and C. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS One*, vol. 6, no. 12, p. e26752, 2011.
11. J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *WSDM*, 2011, pp. 177–186.
12. R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *ICWSM*, 2012.
13. J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential Twitterers," in *WSDM*, 2010, pp. 261–270.
14. C. X. Lin, Q. Mei, Y. Jiang, J. Han, and S. Qi, "Inferring the diffusion and evolution of topics in social communities," *SNMA*, 2011.
15. B. Suh, L. Hong, P. Pirolli, and E. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *SocialCom*, 2010, pp. 177–184.
16. D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *ICWSM 2010*, vol. 5, no. 4, 2010, pp. 130–137.
17. M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *WSDM*, 2013, pp. 23–32.
18. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on Twitter," in *WSDM*, 2011.
19. A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *WSDM*, 2011, pp. 45–54.

20. F. Reali and T. L. Griffiths, "Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift," *Proc. of the Royal Society B: Biological Sciences*, vol. 277, no. 1680, pp. 429–436, 2010.
21. M. De Choudhury, "Tie formation on Twitter: Homophily and structure of egocentric networks," in *PASSAT and SocialCom*. IEEE, 2011, pp. 465–470.
22. J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in Twitter," in *WWW*. ACM, 2012, pp. 251–260.
23. R. Ribiero, "25 small-business Twitter hashtags to follow," http://www.biztechmagazine.com/article/2012/06/25-small-business-twitter-hashtags-follow, 2012.
24. A. K. McCallum, "MALLET: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.
25. J. Rennie, "20 newsgroups," http://www.qwone.com/ jason/20Newsgroups/, 2008.
26. A. Gulli, "News space," http://www.di.unipi.it/ gulli/, 2012.
27. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
28. S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, 1998.
29. E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *WWW*, 2012, pp. 519–528.
30. A. Meyerson and B. Tagiku, "Minimizing average shortest path distances via shortcut edge addition," in *APPROX/RANDOM*, 2009, pp. 272–285.