Soc. Netw. Anal. Min. (2017)7:52
DOI 10.1007/s13278-017-0467-9

ORIGINAL ARTICLE

Defining and evaluating Twitter influence metrics: a higher-order approach in Neo4j

Georgios Drakopoulos[1] · **Andreas Kanavos**[2] · **Phivos Mylonas**[1] · **Spyros Sioutas**[1]

Received: 27 January 2017 / Revised: 23 September 2017 / Accepted: 28 September 2017

Abstract Ranking account influence constitutes an important challenge in social media analysis. Until recently, influence ranking relied solely on the structural properties of the underlying social graph, in particular on connectivity patterns. Currently, there has been a notable shift to the next logical step where network functionality is taken into account, as online social media such as Reddit, Instagram, and Twitter are renowned primarily for their functionality. However, contrary to structural rankings, functional ones are bound to be network-specific since each social platform offers unique interaction possibilities. This article examines seven first-order influence metrics for Twitter, defines a strategy for deriving their higher-order counterparts, and outlines a probabilistic evaluation framework. Experiments with a Twitter subgraph with ground truth influential accounts indicate that a single metric combining structural and functional features outperforms the rest in said framework.

✉ Andreas Kanavos
kanavos@ceid.upatras.gr

Georgios Drakopoulos
c16drak@ionio.gr

Phivos Mylonas
fmylonas@ionio.gr

Spyros Sioutas
sioutas@ionio.gr

[1] Department of Informatics, Ionian University, 49100 Corfu, Greece

[2] Computer Engineering and Informatics Department, University of Patras, 26504 Patras, Greece

Keywords Humanistic data · Higher-order data · Higher-order moments · Influence metrics · Structural metrics · Functional metrics · Twitter · Neo4j

Mathematics Subject Classification 05C82 · 05C85 · 62P25 · 91D30 · 97K30

1 Introduction

Social media reflect underlying social dynamics. The numerous examples make for some rather convincing cases. LinkedIn often determines who shall fill a job vacancy. During the Arab Spring of 2011, Egyptian protesters would communicate in Twitter (Lotan et al. 2011; Tong et al. 2012). In July 2016, the Turkish premier issued a dramatic public address in Skype while a coup was in progress—and failing. Besides the advantage for marketers and political campaign planners among others, social media enable the investigation on a scale considered prohibitive of questions such as social coherence (Leskovec 2011), social graph partitioning (Leskovec et al. 2014), expansion potential (Russell 2013), stable evolutionary strategies (Dawkins 2006), or meme diffusion (Blackmore 2000).

Determining influential accounts is paramount in social network analysis. Currently, the majority of influence metrics relies either on structural properties of the social graph itself (Katz 1953) or on the spectral properties of the associated adjacency matrix (Benzi and Boito 2010; Drakopoulos 2016; Drakopoulos et al. 2015; Fiedler 1973). Prime examples of the former are respectively the number of neighbors and the decay rate of the graph eigenvalues. Both metrics are generic enough to be applied to virtually any social graph. However, they are oblivious to

Published online: 17 October 2017

functionality, a severe limitation as social networks were precisely set up in order to perform specific tasks. Facebook is well known for the *like* button, in Twitter accounts *follow* each other, Foursquare is essentially a closely interwoven fabric of *check-ins*, and *social login* is supported across diverse Web portals as an authentication scheme.

Valuable insight is obtained by harvesting information from Twitter data, namely tweets, hashtags, follows, and mentions. Due to the connection-oriented nature of social data, Neo4j (Panzarino 2014; Robinson et al. 2013) was selected. It is intended to provide reliable and scalable graph storage and, potentially, graph libraries such as NetworkX,[1] Google Pregel (Malewicz et al. 2010), and Spark GraphX[2] can run on top of it.

The primary contribution of this article is twofold. First, a linear algebraic strategy for deriving higher-order Twitter influence rankings from first-order ones is considered, extending these proposed in Drakopoulos et al. (2016b). Second, in order to evaluate their performance, these metrics are compared in a higher-order probablistic framework. Specifically, the first-order metrics proposed in Kafeza et al. (2013, 2014) are extended to higher-order ones by taking into consideration the interaction of the accounts within the particular social graph using techniques from Drakopoulos et al. (2016b).

This article is structured as follows. Section 2 summarizes work in digital influence. The influence metrics and the evaluation framework are outlined in Sect. 3. Section 4 interprets the results, while Sect. 5 examines performance aspects. The article concludes with the discussion in Sect. 6. Article notation is summarized in Table 1. Finally, given the number of institutions and entities maintaining a strong Twitter presence, it makes sense to write about Twitter *accounts* rather than *users*.

Table 1 Article notation

Symbol	Meaning
$\triangleq$	Definition or equality by definition
$\{x_1, \ldots, x_n\}$	Set containing elements $x_1, \ldots, x_n$
$\lvert S \rvert$	Cardinality of set S
$S_1 \setminus S_2$	Asymmetric set difference S_1 minus S_2
τ_{S_1,S_2}	Tanimoto similarity coefficient of sets S_1 and S_2
ν_{S_1,S_2}	Asymmetric Tversky index for sets S_1 and S_2
$\rho_{\mathbf{x}_1,\mathbf{x}_2}$	Correlation coefficient of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$
$\mathrm{E}[X]$	Mean value of random variable X
$\mathrm{Var}[X]$	Variance of random variable X
$\kappa_3(p)$	Pearson skewness coefficient of distribution p
$\kappa_4(q)$	Pearson kurtosis coeffcient of distribution p
$\langle p \lvert\rvert q \rangle$	Kullback–Leibler divergence between distributions p and q
$\mu^1 \succ \mu^2$	Metric μ^1 always outperforms μ^2
$\mu^1 \succeq \mu^2$	Metric μ^1 is at least as good as μ^2
$\mathbf{1}_n$	$n \times 1$ vector of ones
$\varphi(\cdot)$	Sigmoid or logistic function
$\mathrm{sgn}(\cdot)$	Sign function

2 Related work

Quoting (Cha et al. 2010) social influence is

> the ability of a person to influence the thoughts or actions of others.

Well before the advent of social media, scientific literature was already abounding with influence metrics based either on tangible factors, such as total income, or on abstract concepts, like the quadruple of the Roman values of *gravitas*, *pietas*, *dignitas*, and *virtus* (Manicas 1991; Turner 1991; Zimbardo and Leippe 1991). In Katz (1953) a combinatorial metric based on connectivity patterns of the underlying social graph is proposed. Similar rankings include betweeness- (Newman 2005), degree- (Bonacich 1987), and closeness centrality (Okamoto et al. 2008). Algebraic metrics include the eigenvector centrality (Drakopoulos et al. 2015), the Estrada index (Li et al. 2009), and the matrix power series (Drakopoulos and Megalooikonomou 2016; Estrada and Higham 2010; Rivest and Vuillemin 1976). For the relationship between combinatorial and algebraic rankings, both of structural nature, see Benzi and Klymko (2015).

Functional metrics, as their name suggests, focus on the functionality of a social network and, consequently, facilitate interpretation at the expense of universal applicability. Regarding Twitter, personality models have been used for community discovery (Kafeza et al. 2013, 2014), probabilistic analysis predict the most trending authors for a given topic (2011 a). Concerning the digital influence of a Twitter account, it can be derived by PageRank extensions (Weng et al. 2010; TunkRank 2015), by its importance compared to that of the remaining network (Mehta et al. 2012), or by a nonlinear combination of features[3] (Razis and Anagnostopoulos 2014). Real-time influence analytics were proposed in Zamparas et al. (2015). The discrimination between authoritative and non-authoritative accounts in Yahoo! Answers can be achieved through modeling authority as a mixture of gamma distributions (Bouguessa et al. 2008). Influence is examined in light of the current technological evolution, which eventually led to the

[1] https://networkx.github.io.

[2] http://spark.apache.org/graphx.

[3] http://www.influencetracker.com.

Table 2 Data for the kth account

Feature	Meaning	Feature	Meaning	Feature	Meaning
T_k	Tweet set	Φ_k	Follower set	C_k	Reply set
R_k	Retweet set	Ψ_k	Followee set	M_k	Mention set
H_k	Hashtag set	V_k	Favorites set	F_k	Frequency

creation of social media (Rogers and Beal 1957). Under plausible assumptions the most influential accounts are the cost effective, where cost is a function of communication complexity (Bakshy et al. 2011).

The same ranking may yield different results to the same set of accounts across social media (Smith et al. 2012). Moreover, the border between structural and functional may not always be clear as in the case of PageRank (Page et al. 1999). Moving beyond the structural and functional metric distinction, fusion strategies for creating improved influence rankings based on tensor algebra are proposed in Drakopoulos (2016) and Drakopoulos and Kanavos (2016).

More recently, the issue of online trust has been tied to that of influence. In Golbeck (2009) influence is mainly a matter of trust. Signed networks for modeling account trustworthiness have been proposed in Leskovec et al. (2010). Alternatively, agents for collecting and evaluating trust-related data are designed in Bickmore and Cassell (2001). Features for trusting players in networks of online gamers are extracted in Gao (2005). Online news validity is augmented with account trustworthiness in Bodnar et al. (2014). In Bertot et al. (2010) social media are regarded as pylons of an open society and of government accountability.

3 Twitter influence metrics

In order to differentiate between the existing metrics of Kafeza et al. (2013, 2014) and the proposed ones, the following definitions are necessary.

Definition 1 First-order metrics compute the digital influence of any account based only on data concerning this account.

Definition 2 Higher-order metrics derive the digital influence of any account as a function of the influence of other accounts. Consequently, the data regarding a specific account alone are insufficient for computing its digital influence.

3.1 First-order metrics

Seven first-order Twitter influence rankings and their formulation as Cypher queries are overviewed. Six have already been proposed Drakopoulos et al. (2016b), Kafeza et al. (2013, 2014) and Kanavos et al. (2014a, b), whereas the last one is new. Table 2 summarizes the features, which can be either directly collected from Twitter or computed by these rankings.

The difference between C_k and M_k is that the @ handle is respectively at the beginning and anywhere but the beginning of the tweet. Frequency F_k is defined as the sum of tweets and retweets every eight hours.

First-order metric 1 *Conversational accounts have a high number of tweets, retweets, conversations, favorites, and mentions. Thus, they relay a significant amount of information. The conversational metric μ^c is calculated as*

$$\mu_k^c \triangleq |T_k| + |R_k| + |C_k| + |V_k| + |M_k|. \tag{1}$$

First-order metric 2 *Multisystemic accounts have a high number of hashtags in their tweets, retweets, and conversations. These accounts are probably proficient in a broad range of topics and they are a likely point of reference. The multisystemic metric is denoted by μ^m and is calculated as follows*

$$\mu_k^m \triangleq |H_k|. \tag{2}$$

First-order metric 3 *Active accounts have a high number of tweets over a given time interval. This behavior pattern likely indicates knowledge of or strongly opinion about a particular topic. Thus, anyone seeking to know about this topic might consider an active account an authority. The account/energetic metric is denoted by μ^e and is calculated as*

$$\mu_k^e \triangleq F_k. \tag{3}$$

First-order metric 4 *Popular accounts have a high number of followers. Although Twitter popularity does not necessarily correspond to optimal diffusion, highly followed users exert limited influence since they are often read. The popularity metric μ^p is computed as*

$$\mu_k^p \triangleq |\Phi_k|. \tag{4}$$

First-order metric 5 *Active accounts maintain a relative balance between the sets of accounts that follow and the accounts who are followed by them. This is expressed by the Tanimoto similarity coefficient for sets Φ_k and Ψ_k. In this case, ranking μ^a is computed as*

$$\mu_k^a \triangleq \tau_{\Phi_k,\Psi_k} = \frac{|\Phi_k \cap \Psi_k|}{|\Phi_k \cup \Psi_k|} = \frac{|\Phi_k \cap \Psi_k|}{|\Phi_k| + |\Psi_k| - |\Phi_k \cap \Psi_k|}. \tag{5}$$

Notice that the second form is computationally appealing compared to the first, as $|\Psi_k|$ and $|\Phi_k|$ are readily available, while set intersection queries typically return fewer items than the corresponding union queries.

First-order metric 6 *Another way to define active accounts is to consider the ranking μ^s which relies on the sigmoid function $\varphi(\cdot)$*

$$\mu_k^s \triangleq \varphi(s) = \frac{1}{1+e^{-s}}, \quad s \triangleq \frac{\log(1+|\Phi_k|)}{1+\log(1+|\Psi_k|)}. \tag{6}$$

First-order metric 7 *The atomic influential metric, denoted by μ^i, computes the geometric mean of many of the above features*

$$\mu_k^i \triangleq (|T_k||R_k||H_k| \log_{10}(1+|\Phi_k|))^{\frac{1}{4}} \tag{7}$$

in order to capture the total online presence of an account.

3.2 Higher-order metrics

Since graphs are primarily about connectivity and interaction, it makes sense to seek influence rankings which fall under Definition 2. Such is the case of Katz centrality (Katz 1953) which relies on the directed adjacency matrix $\tilde{\mathbf{A}}$ of the social graph of acquaintances among n individuals to compute the score

$$\mu_k^Z \triangleq \sum_{p=1}^{+\infty} \sum_{j=1}^{n} \alpha_0^p \tilde{\mathbf{A}}^p[j,k]. \tag{8}$$

By means of the Neumann identity and provided that the spectral radius of $\alpha_0\tilde{\mathbf{A}}$, the largest in absolute value eigenvalue is strictly less than one

$$\sum_{p=0}^{+\infty} (\alpha_0\tilde{\mathbf{A}})^p = (\mathbf{I}_n - \alpha_0\tilde{\mathbf{A}})^{-1}; \tag{9}$$

the score formula can be recast as the matrix equation

$$\mu^{\mathbf{Z}} = \left(\left(\mathbf{I}_n - \alpha_0\tilde{\mathbf{A}}^T\right)^{-1} - \mathbf{I}_n \right)\mathbf{1}_n = \left(\mathbf{I}_n - \alpha_0\tilde{\mathbf{A}}^T\right)^{-1}\mathbf{1}_n - \mathbf{1}_n. \tag{10}$$

Within the context of the analysis by Katz, the meaning of the direction, namely the fact that $\tilde{\mathbf{A}}$ may in the general case be non-symmetric, is that a person may know another indirectly, for instance through rumors or by a random mention from mutual acquaintances, whereas the converse need not be true. $\tilde{\mathbf{A}}$ expresses who knows who, while $\tilde{\mathbf{A}}^T$ represents who is known by whom and, therefore, is a metric of fame.

Along similar lines, the TunkRank algorithm (TunkRank 2015) is a functional higher-order influence ranking designed for Twitter. For a given account the TunkRank metric μ^R connects the set of followers Φ_k, the number $|\Psi_j|$ of their followers, and the network average retweet probability p_0 in the succinct formula

$$\mu_k^R \triangleq \sum_{j\in\Phi_k} \frac{1+p_0\,\mu_j^R}{|\Psi_j|}, \quad 0 \le p_0 \le 1 \tag{11}$$

which essentially states that μ_k^R is a linear combination of that of its followers which in turn is another linear combination of their respective followers and so forth. If μ_k^R are stacked to a column vector $\mu^{\mathbf{R}}$, then

$$\mu^{\mathbf{R}} = \mathbf{B}\mu^{\mathbf{R}} \Leftrightarrow (\mathbf{I}_n - \mathbf{B})\mu^{\mathbf{R}} = \mathbf{0}, \quad \mathbf{B} \in \mathbb{R}^{n\times n} \tag{12}$$

which, like PageRank, can be cast either as a linear system or as an eigenvector problem. In either case the solution lies in the non-trivial nullspace of $\mathbf{I}_n - \mathbf{B}$.

Because of their popularity in literature and also because of the fact that the former is a purely structural metric whereas the latter is solely functional, the Katz and the TunkRank were selected as baselines.

The general scheme proposed in this article to construct a higher-order influence ranking from a first-order one consists of two phases. First, each original score μ_k' is normalized to μ_k in the range $[\vartheta_0, 1]$ with the transform

$$\mu_k = \max\left\{ \vartheta_0, \frac{\mu_k' - \min_{1\le j\le n}\left\{\mu_j'\right\}}{\max_{1\le j\le n}\left\{\mu_j'\right\} - \min_{1\le j\le n}\left\{\mu_j'\right\}} \right\}, \quad 1 \le k \le n \tag{13}$$

where ϑ_0 is an arbitrarily chosen lower bound.

During the second phase, the *authority score* of each account is computed as the convex combination of its own score and that of its followers

$$\mu_k^{\text{auth}} = \eta_0\mu_k + \frac{(1-\eta_0)}{|\Phi_k|}\sum_{j\in\Phi_k}\mu_j, \quad 0<\eta_0<1 \tag{14}$$

Stacking $\mu_k, \mu_k^{\text{auth}}$ to vectors $\mu, \mu^{\text{auth}} \in \mathbb{R}^n$ leads to the matrix equation

$$\mu^{\text{auth}} = \eta_0\mu + (1-\eta_0)\mathbf{M}\mathbf{1}_n \tag{15}$$

where the matrix $\mathbf{M}$ is defined elementwise as

$$\mathbf{M}[i,j] \triangleq \begin{cases} \dfrac{\mu_j}{|\Phi_i|}, & j \in \Phi_i \\ 0, & j \notin \Phi_i \end{cases} \in \mathbb{R}^{n\times n}. \tag{16}$$

Likewise the *hub score* is computed using the followed accounts

$$\mu_k^{\text{hub}} = \eta_1 \mu_k + \frac{(1-\eta_1)}{|\Psi_k|} \sum_{j \in \Psi_k} \mu_j, \quad 0 < \eta_1 < 1 \tag{17}$$

resulting in the matrix equation

$$\mu^{\text{hub}} = \eta_1 \mu + (1-\eta_1)\mathbf{M}^T \mathbf{1}_n. \tag{18}$$

Finally, the F1 metric of the authority and the hub scores is computed

$$\mu_k^{F1}(\eta_2) = \frac{1+\eta_2}{\frac{1}{\mu_k^{\text{auth}}} + \frac{\eta_2}{\mu_k^{\text{hub}}}} = \frac{(1+\eta_2)\mu_k^{\text{auth}}\mu_k^{\text{hub}}}{\mu_k^{\text{hub}} + \eta_2 \mu_k^{\text{auth}}}. \tag{19}$$

Note that the F1 metric is frequently employed in information retrieval problems as it systematically provides a deeper insight than the precision P or recall R scores alone. It is defined as the weighted harmonic mean of P and R

$$F1(P,R;\gamma_0) \triangleq \frac{1+\gamma_0}{\frac{1}{P} + \frac{\gamma_0}{R}} = (1+\gamma_0)\frac{PR}{R+\gamma_0 P}. \tag{20}$$

In this way from any of the preceding first-order influence metrics, a higher-order ranking can be derived. As a convention, the capital letter of the corresponding small letter denoting a first-order metric will be used. Thus, μ^M and μ^A are the higher-order counterparts of μ^m and μ^a, respectively.

4 Results

4.1 Data synopsis and baseline ranking

In order to evaluate the influence metrics, a Twitter subgraph G_b was collected during November and December of 2016. As in Drakopoulos et al. (2016b) the starting point of the social crawler, programmed to move along follow relationships to accounts tweeting educational hashtags, was the official Twitter of a major US university. The vertices of G_b are Twitter accounts and the edges indicate following relationships. Its properties are stated in Table 3 and indicate a Twitter subnetwork that displays considerable activity. For the definitions of density and completeness, see Drakopoulos et al. (2016b).

Table 3 Structural (right) and functional (left) properties of G_b

Property	Value	Property	Value
Vertices	12,731	Distinct hashtags	739
Edges	238,992	Hashtags	18,221
Triangles	4364	Tweets	21,217
Squares	471	Retweets	13,445
Density	18.77	Average following	4.33
Completeness	0.0029	Average followers	7.61
Diameter	37	Average tweet length	87.11

By enlisting the aid of a domain expert, the influential accounts shown in Table 4 were identified in G_b, the number of which corroborates the high activity in G_b as well as the broad flexibility any ranking scheme has on this dataset, as approximately 16% of the accounts are influential. For the purposes of this analysis, μ^Z and μ^R were the baseline metrics on the grounds that their Zipf exponents, defined in Eqs. (23) and (24), were the closest to the exponents of the ranking derived by the expert. Since μ^Z is structural whereas μ^R is functional, it is of interest to determine which rankings, if any, are closer to these baselines. The reason for selecting baseline metrics is that an expert may not be available or there are way too many accounts to rank for a single human or even for a group of humans.

The criteria the domain expert was based on were the following:

- Institutions and organizations are more influential compared to individuals.
- Official accounts are reference points in the Web.
- In academia, faculty members are traditionally treated with respect.
- Sports associations are mainstays of US academic life.

Let S be the set of accounts in G_b. By partitioning the rankings obtained by each metric to b bins where

$$b = 1 + \left\lceil \sqrt{|S|} \right\rceil, \quad b \equiv 0 (\text{mod } 2) \tag{21}$$

and by counting the number of influential accounts $|S_k|$ in the kth bin divided by $|S|$, then each ranking can be mapped to a discrete distribution. Let p^* be the distribution corresponding to μ^*. Therefore,

$$p_k^* = \frac{|S_k|}{\sum_{k=1}^{b} |S_k|} = \frac{|S_k|}{|S|}, \quad 1 \le k \le b. \tag{22}$$

Then, the quality of each influence metric can be assessed within a probabilistic framework such as the one outlined in the next subsections. The selection of b as in Eq. (21) keeps both the estimation complexity and the variance low.

Table 4 Breakdown of known influential accounts in G_b

Account type	Number
Student organizations	693
Faculty members	471
Student leaders	397
Sports	306
Other academia	118
Departments	98
Schools	61
Universities	17

To facilitate the analysis, b was intentionally chosen to be even. Also, as a convention, the distribution will retain the letter denoting the corresponding metric and, thus, for instance p^S and p^m correspond to μ^S and μ^m, respectively.

For the final evaluation of each metric, the following two definitions from Drakopoulos et al. (2016b) will be used.

Definition 3 Assume a fixed set T of influence metric evaluation tests. Metric μ^1 outperforms μ^2 with respect to T if and only if μ^1 achieves strictly better evaluation scores than μ^2 in each test of T. This case is denoted as $\mu^1 \succ \mu^2$.

Definition 4 Assume a fixed set T of influence metric evaluation tests. Metric μ^1 is at least as good as μ^2 with respect to T if and only if μ^1 achieves

- strictly better evaluation score than μ^2 in at least one test of T
- the same evaluation score with μ^2 in the remaining tests.

This case is denoted as $\mu^1 \succeq \mu^2$.

The parameters for deriving the higher-order rankings were $\eta_0 = 0.4$, $\eta_1 = 0.4$, and $\eta_2 = 0.6$. These values of η_0 and η_1 imply that the network effect in shaping the authority and the hub scores, respectively, is 20% larger compared to the ranking value computed for each account by the first-order metric. This is in accordance to the spirit of higher-order metrics, which are by construction connection-oriented. The value of η_2 means that the authority score contributes 20% more than the hub score to the final influence ranking.

4.2 Architecture

Figure 1a, b illustrates the components as well as the information flow between them in the open-loop architecture of Drakopoulos et al. (2016b). First the social crawler populates the database and then analysis follows. The social crawler has been written in Python using Tweepy[4] for collecting non-streaming Twitter data. Neo4j version 3.0 was configured in embedded mode, meaning that a single JVM was launched. Thus, system memory was conserved at the expense of making the database accessible only to the client (Robinson et al. 2013).

4.3 Ranking score clustering

Substantial evidence suggests that influence metric scores of large graphs show a strong tendency to be clustered according to a Zipf model (Leskovec et al. 2014)

$$p_k = \alpha_0 k^{-\gamma_0}, \quad \alpha_0, \gamma_0 > 0 \tag{23}$$

or to a cutoff Zipf model (Russell 2013)

$$p_k = \alpha_1 k^{-\gamma_1} e^{-\beta_1 k}, \quad \alpha_1, \beta_1, \gamma_1 > 0. \tag{24}$$

Moreover, a sorted version of the DGX distribution (Bi et al. 2001) which includes Eq. (23) as a special case was used as a ranking model

$$\begin{aligned} \text{prob}\{X_{\text{dgx}} = k\} &= \frac{1}{\beta_0 k}\exp\left(-\frac{(\ln k - \mu_0)^2}{2\sigma_0^2}\right), \beta_0 > 0 \\ \beta_0 &= \sum_{k=1}^{b}\frac{1}{k}\exp\left(-\frac{(\ln k - \mu_0)^2}{2\sigma_0^2}\right). \end{aligned} \tag{25}$$

When sorted, the DGX distribution has an initial steep decay followed by a tail which is heavier than that of the Gaussian distribution but also decays quicker than a Zipf distribution.

Therefore, comparison any of p and p^* can be reduced to the distance between their respective scalar parameters, namely γ_0 for Eq. (23), γ_1 and β_1 for Eq. (24), and μ_0 and σ_0^2 for Eq. (25).

Linearizing Eqs. (23) and (24) for each point and stacking the equations row-wise as in Drakopoulos et al. (2016b) yields respectively the overdetermined linear systems

$$\begin{bmatrix} \ln p_1 \\ \ln p_2 \\ \vdots \\ \ln p_b \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\ln 2 \\ \vdots & \vdots \\ 1 & -\ln b \end{bmatrix} \begin{bmatrix} \ln \alpha_0 \\ \gamma_0 \end{bmatrix} \quad \text{and}$$

$$\begin{bmatrix} \ln p_1 \\ \ln p_2 \\ \vdots \\ \ln p_b \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -2 & -\ln 2 \\ \vdots & \vdots & \vdots \\ 1 & -b & -\ln b \end{bmatrix} \begin{bmatrix} \ln \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix}. \tag{26}$$

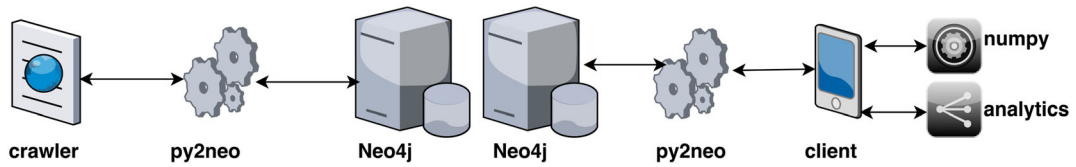
The normal systems for the above cases are respectively

$$\underbrace{\begin{bmatrix} \sum_{k=1}^{b} \ln p_k \\ -\sum_{k=1}^{b} \ln p_k \ln k \end{bmatrix}}_{\mathbf{v}_2} = \underbrace{\begin{bmatrix} b & -\sum_{k=1}^{b} \ln k \\ -\sum_{k=1}^{b} \ln k & \sum_{k=1}^{b} \ln^2 k \end{bmatrix}}_{\Sigma_2} \begin{bmatrix} \ln \alpha_0 \\ \gamma_0 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \sum_{k=1}^{b} \ln p_k \\ -\sum_{k=1}^{b} k \ln p_k \\ -\sum_{k=1}^{b} \ln p_k \ln k \end{bmatrix}}_{\mathbf{v}_3} = \underbrace{\begin{bmatrix} b & \frac{b(b-1)}{2} & -\sum_{k=1}^{b} \ln k \\ \frac{b(b-1)}{2} & \frac{b(b+1)(2b+1)}{6} & \sum_{k=1}^{b} k \ln k \\ -\sum_{k=1}^{b} \ln k & \sum_{k=1}^{b} k \ln k & \sum_{k=1}^{b} \ln^2 k \end{bmatrix}}_{\Sigma_3} \begin{bmatrix} \ln \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix}. \tag{27}$$

Ignoring the estimators for normalizing constants $\hat{\alpha}_0$ and $\hat{\alpha}_1$, which are nuisance parameters, the least squares estimators $\hat{\gamma}_0$, $\hat{\beta}_1$, and $\hat{\gamma}_1$ are

[4] http://www.tweepy.org.



(a) Twitter subgraph creation. **(b)** Analytics operation and data flow.

Fig. 1 System architecture

$$\begin{aligned}
\hat{\gamma}_0 &= \frac{\det\left(\Sigma_2^{(2)};\mathbf{v}_2\right)}{\det(\Sigma_2)} \\
&= -\frac{b\left(\sum_{k=1}^{b}\ln p_k \ln k\right) - \left(\sum_{k=1}^{b}\ln p_k\right)\left(\sum_{k=1}^{b}\ln k\right)}{b\left(\sum_{k=1}^{b}\ln^2 k\right) - \left(\sum_{k=1}^{b}\ln k\right)^2} \\
\hat{\gamma}_1 &= \frac{\det\left(\Sigma_3^{(3)};\mathbf{v}_3\right)}{\det(\Sigma_3)}, \quad \det(\Sigma_3) \neq 0 \\
\hat{\beta}_1 &= \frac{\det\left(\Sigma_3^{(2)};\mathbf{v}_3\right)}{\det(\Sigma_3)}, \quad \det(\Sigma_3) \neq 0
\end{aligned} \tag{28}$$

where $\Sigma_j^{(i)};\mathbf{v}_j$ is the matrix resulting by replacing the ith column of Σ_j with $\mathbf{v}_j$. The determinants of 3×3 matrices can be symbolically computed with the rule of Sarrus, which is a special case of the rule of Leibniz. The finite sums in Eq. (27) involving logarithms can be approximated for large b by the Euler–McLaurin summation formula

$$\begin{aligned}
\sum_{k=1}^{b}\ln^n k \approx\ & b\left(\ln^n b - n\ln^{n-1} b + n(n-1)\ln^{n-2} b - \cdots + (-1)^n n!\right) \\
& + \frac{1}{2}\ln^n b + \frac{n}{12b}\ln^{n-1} b + \xi_0 + \mathrm{O}\left(\frac{\ln^{n-1} b}{b^3}\right), \quad \xi_0 \in \mathbb{R}.
\end{aligned} \tag{29}$$

When b is large, then alternatively the following approximations may be used when formulating the systems in Eq. (27)

$$\begin{aligned}
\sum_{k=1}^{b} k^n &\approx \int_1^b x^n \mathrm{d}x + \xi_1 = \frac{b^{n+1}-1}{n+1} + \xi_1, \quad \xi_1 \in \mathbb{R} \\
\sum_{k=1}^{b}\ln k &\approx \int_1^b \ln x \mathrm{d}x + \xi_2 = (b-1)\ln b + \xi_2, \quad \xi_2 \in \mathbb{R}
\end{aligned} \tag{30}$$

where ξ_1 and ξ_2 are correction constants.

The optimal values for μ_0 and σ_0^2 can be determined by numerically optimizing with a procedure from Bi et al. (2001) the loglikelihood function

$$\ell(\mu_0,\sigma_0) = b\ln\beta_0 - \sum_{k=1}^{b}\left(-\ln p_k + \frac{(\ln p_k - \mu_0)^2}{2\sigma_0^2}\right). \tag{31}$$

Table 5 Differences between μ^Z (left) and μ^R (right) and the reference ranking

	$\lvert\gamma_0^* - \hat{\gamma_0}\rvert$	$\lvert\gamma_1^* - \hat{\gamma_1}\rvert$		$\lvert\gamma_0^* - \hat{\gamma_0}\rvert$	$\lvert\gamma_1^* - \hat{\gamma_1}\rvert$
μ^Z	0.4991	0.5513	μ^R	0.4223	0.5823

The absolute differences of the exponents between μ^Z and μ^R and the ranking provided by the expert are shown in Table 5, while the absolute differences for each estimated parameter are shown in Table 6. We consider the exponents γ_0^* and γ_1^* of the ranking provided by the expert as the Zipf exponents of the true ranking.

4.4 Correlation With p^*

A common metric in probability theory for determining the similarity of any two vectors is the normalized correlation coefficient, which is defined as

$$\rho_{p,p^*} \triangleq \frac{\sum_{k=1}^{b} p_k p_k^*}{\left(\sum_{k=1}^{b} p_k^2\right)^{\frac{1}{2}}\left(\sum_{k=1}^{b}\left(p_k^*\right)^2\right)^{\frac{1}{2}}} = \frac{\mathbf{p}^T\mathbf{p}^*}{||\mathbf{p}||_2||\mathbf{p}^*||_2}. \tag{32}$$

The second form of Eq. (32) has the geometric interpretation that ρ_{p,p^*} is the cosine of the inner angle θ_0 formed by $\mathbf{p}$ and $\mathbf{p}^*$ in $\mathbb{R}^b$, that is

$$\theta_0 \triangleq \arccos\left(\rho_{p,p^*}\right), \quad \theta_0 \in \left[0, \frac{\pi}{2}\right]. \tag{33}$$

In this case $\rho_{p,p^*} \geq 0$ since rankings are positive as a result of Eq. (13). Moreover, $\rho_{p,p^*} \leq 1$ because of the Cauchy–Schwartz inequality. Since both $\mathbf{p}$ and $\mathbf{p}^*$ are both positive, θ_0 can belong only to the first quadrant. Table 7 shows the correlation between the metrics and the baselines as well as θ_0 in degrees.

4.5 Kullback–Leibler divergence from p^*

A common divergence metric for measuring the distance between two distributions in terms of information theory is

Table 6 Parameter differences for the Katz and the TunkRank baselines

μ^Z	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
$\lvert\hat{\gamma}_0 - \hat{\gamma}_0^*\rvert$	1.1989	1.0223	1.9055	1.6228	1.4000	1.3000	0.9855	1.0192
$\lvert\hat{\gamma}_1 - \hat{\gamma}_1^*\rvert$	1.1313	0.9922	1.5175	1.3981	1.3114	1.2200	0.7022	1.0899
$\lvert\hat{\beta}_1 - \hat{\beta}_1^*\rvert$	1.6663	1.4991	3.0022	2.5463	2.2878	1.8890	1.2286	1.3316
$\lvert\hat{\mu}_0 - \hat{\mu}_0^*\rvert$	1.6618	1.5512	2.6845	2.2046	2.0332	1.8753	1.3113	1.4824
$\lvert\hat{\sigma}_0^2 - \hat{\sigma}_0^{2*}\rvert$	1.1742	0.9916	2.1003	1.9000	1.7000	1.3000	0.8236	0.8536
μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
$\lvert\hat{\gamma}_0 - \hat{\gamma}_0^*\rvert$	1.4529	1.3783	1.8873	1.7842	1.6483	1.1021	0.8916	0.9612
$\lvert\hat{\gamma}_1 - \hat{\gamma}_1^*\rvert$	1.2751	1.1222	1.4031	1.3529	1.3011	0.9981	0.6992	0.8733
$\lvert\hat{\beta}_1 - \hat{\beta}_1^*\rvert$	1.7025	1.5878	2.4767	2.2741	1.9442	1.2503	1.0044	1.1296
$\lvert\hat{\mu}_0 - \hat{\mu}_0^*\rvert$	1.5773	1.3742	2.5590	2.1331	1.7462	1.1998	0.8523	1.1278
$\lvert\hat{\sigma}_0^2 - \hat{\sigma}_0^{2*}\rvert$	1.6216	1.4332	2.2331	2.1533	2.0023	1.2001	0.9125	1.1161

Table 7 Correlation with the Katz and the TunkRank baselines

ρ_{p,p^*}	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
μ^Z	0.5931	0.5933	0.3271	0.3269	0.4551	0.5102	0.6813	0.6619
μ^R	0.3903	0.4212	0.1844	0.2354	0.2341	0.6845	0.7201	0.6930
θ_0	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
μ^Z	53.62	53.60	70.90	70.91	62.92	59.32	47.05	48.55
μ^R	67.02	65.08	79.37	76.38	76.46	46.80	43.93	46.13

Table 8 Divergence between metrics and the Katz and the TunkRank baselines

$\langle p\|\|p^*\rangle$	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
μ^Z	0.5887	0.6219	0.4418	0.4757	0.5423	0.5411	0.6911	0.6743
μ^R	0.5108	0.5111	0.4147	0.4374	0.4572	0.5113	0.5583	0.5220

the Kullback–Leibler divergence. The general formula applied to $\mathbf{p}$ and $\mathbf{p}^*$ yields

$$\langle p||p^*\rangle \triangleq \sum_{k=1}^{b} p_k \log\left(\frac{p_k}{p_k^*}\right) = \sum_{k=1}^{b} p_k \log p_k - \sum_{k=1}^{b} p_k \log p_k^* \tag{34}$$

and essentially is the cross-entropy of $\mathbf{p}$ and $\mathbf{p}^*$ minus the entropy of $\mathbf{p}^*$. Note that $\mathbf{p}$ and $\mathbf{p}^*$ are not interchangeable. This happens because, unlike the normalized correlation coefficient, the Kullback–Leibler divergence can distinguish between a reference distribution, $\mathbf{p}^*$ in this case, and a variant, namely $\mathbf{p}$. Table 8 shows the values of $\langle p||p^*\rangle$ for each pair of influence rankings.

4.6 Tversky divergence from p^*

Another asymmetric divergence metric for measuring the pairwise distance between p and p^* is the Tversky index. The latter is defined for two sets T, which is considered as a template, and V, the variant, as

$$v_{T,V} \triangleq \frac{|T \cap V|}{|T \cap V| + w_1|T \setminus V| + w_2|V \setminus T|} \in [0,1], \quad w_1, w_2 > 0 \tag{35}$$

Divergence is inversely proportional to $v_{T,V}$. Hence, a value of 1 denotes full coincidence of the two sets, whereas a value of 0 implies there is no similarity at all. Typically, $w_1 > w_2$ since an element of T not present in V is a distortion of the template set. When both weights equal 1, the Tanimoto coefficient is obtained. Moreover, when $w_1 + w_2 = 1$, then the minimum distance between T and V is maximized. For the purposes of this analysis, $w_1 = 2w_2$ and, hence $w_1 = 2/3$ and $w_2 = 1/3$.

The Tversky index was originally designed for sets but it can be used with vectors as well with some modifications. If it is applied directly to $\mathbf{p}$ and $\mathbf{p}^*$ by placing their elements

to T and V, respectively, then the result may equal 1 even when $\mathbf{p}$ and $\mathbf{p}^*$ do not match. For instance, this may happen if they are a permutation of each other. Instead, to avoid this and to obtain higher granularity, the b points of $\mathbf{p}$ and $\mathbf{p}^*$ are partitioned to $s+1$ consecutive segments where the first s ones contain $\lfloor b/s \rfloor$ points and the last one comprises of $b \bmod s$ points. Let T_j and V_j denote the sets containing the elements of these segments where $1 \le k \le s+1$. Then the Tversky index is applied to each pair of T_j and V_j to yield the similarity score

$$v_{s+1} \triangleq \frac{1}{s+1} \sum_{j=1}^{s+1} v_{T_j,V_j} \tag{36}$$

For the purposes of this analysis $s+1 = \lceil b/\log b \rceil$. This selection guarantees that there will be a large number of sets each with few but still enough samples, namely $O(\log b)$, so that reliable values for v_{s+1} can be computed. This contrasts the balance between the number of sets and the samples in each set achieved in Eq. (21) when b was selected. Table 9 summarizes the $v_{\lceil b/\log b \rceil}$ scores.

4.7 Skewness And Kurtosis compared to p^*

The Pearson skewness coefficient of a random variable X is a higher-order index indicating whether the distribution of X is symmetric or has a high mass concentration either to the left or to the right of its expected value $\mathrm{E}[X]$. It is defined as

$$\kappa_3(X) \triangleq \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^3\right]}{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]^{\frac{3}{2}}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^3\right]}{\mathrm{Var}[X]^{\frac{3}{2}}}. \tag{37}$$

In the present analysis, the following approximations are used

$$\begin{aligned} \mathrm{E}[X] &\approx \mu_b = \frac{1}{b}\sum_{k=1}^{b} p_k \\ \mathrm{Var}[X] &\approx \sigma_b^2 = \frac{1}{b-1}\sum_{k=1}^{b} (p_k - \mu_b)^2. \end{aligned} \tag{38}$$

It is expected that the mass distribution of p^* will be uneven as the majority of the true influential accounts will be located in the left part of the distribution toward the origin point. Therefore, $\kappa_3(X)$ is expected to be positive. Since not only the value of the skewness coefficient is important but also its sign, the following two metrics will be used to assess the deviance of p from p^* in terms of symmetry

$$\delta_3(p) \triangleq |\kappa_3(p) - \kappa_3(p^*)| \tag{39}$$

$$\pi_3(p) \triangleq \begin{cases} 1, & \mathrm{sgn}(\kappa_3(p)) = \mathrm{sgn}(\kappa_3(p^*)) \\ 0, & \mathrm{sgn}(\kappa_3(p)) \neq \mathrm{sgn}(\kappa_3(p^*)). \end{cases} \tag{40}$$

Similarly, the Pearson kurtosis coefficient is a higher-order measure of the degree of mass concentration around $\mathrm{E}[X]$. Namely, small values indicate a strong concentration, while large values are often attributed to a slowly decaying curve or to a big number of outliers. The definition is

$$\kappa_4(X) \triangleq \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^4\right]}{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]^2} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^4\right]}{\mathrm{Var}[X]^2}. \tag{41}$$

As $\kappa_4(X)$ can by definition take only positive values, only the following metric can be defined

$$\delta_4(p) \triangleq |\kappa_4(p) - \kappa_4(p^*)|. \tag{42}$$

Table 10 contains the values for $\delta_3(p)$, $\pi_3(p)$, and $\delta_4(p)$.

4.8 Combining metrics

Based on the results of the preceding tests, the following metric ordering can be inferred when the baseline is μ^Z

$$\mu^I \succeq \mu^{F1} \succ \mu^M \succeq \mu^C \succ \mu^S \succeq \mu^A \succ \mu^P \succeq \mu^E \tag{43}$$

which maintains most of the ordering in Drakopoulos et al. (2016b) between the first-order metrics

$$\mu^i \succ \mu^m \succ \mu^c \succ \mu^p \succeq \mu^e. \tag{44}$$

When the reference metric is μ^R, then the resulting ranking ordering is

$$\mu^I \succ \mu^{F1} \succ \mu^S \succeq \mu^M \succeq \mu^C \succ \mu^A \succeq \mu^P \succ \mu^E. \tag{45}$$

In both orderings μ^I is always better than the remaining metrics, while μ^P and μ^E are always dominated by the rest. This can be attributed to the fact that μ^I combines the main functional and the primary structural features of Twitter. On the contrary, it appears that having many followers alone as in μ^P or tweeting a lot during a specified interval as in μ^E is not always a sign of influence.

The middle part of both orderings reveals some interesting relationships. Both parts have only one clear difference and it is almost the same, namely $\mu^C \succ \mu^S$ and

Table 9 Divergence from the Katz and the TunkRank baselines

$v_{\lceil b/\log b \rceil}$	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
μ^Z	0.2791	0.3056	0.1298	0.1318	0.1993	0.2246	0.3337	0.3489
μ^R	0.2812	0.2687	0.1589	0.2231	0.2240	0.3175	0.3498	0.3536

Table 10 Differences from the Katz and the TunkRank baselines

μ^Z	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
$\delta_3(p)$	1.2355	1.1255	2.3321	1.8816	1.7371	1.6368	**0.9635**	1.0993
$\pi_3(p)$	1	1	0	1	1	1	**1**	1
$\delta_4(p)$	1.5272	1.3845	2.4449	2.2594	2.0125	1.8844	**1.1140**	1.3914
μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
$\delta_3(p)$	1.6899	1.5612	2.0018	1.9032	1.7723	1.3347	**1.2261**	1.3285
$\pi_3(p)$	1	1	1	1	1	1	**1**	1
$\delta_4(p)$	1.7993	1.6029	2.2173	2.0992	1.9544	1.4517	**1.3324**	1.5119

Bold values indicate the numbers of μ^I

$\mu^C \succ \mu^A$ -recall that μ^S and μ^A depend on different ways only on the similarities of Φ_k and Ψ_k. These distinctions imply that a fair summary of online activity is probably a better influence indicator than the *follow* relationships, even when their bidirectionality is factored. Also, it seems that μ^S outperforms μ^A, probably because the former operates on orders of magnitude instead on the set cardinalities directly, resulting thus in better numerical properties and allowing easier handling of uneven set sizes. Since μ^I and μ^C seem to be good choices, while μ^E does not, it follows that tweeting must be coupled with retweeting in order to yield a more reliable indicator, meaning that an influential account should not only post reliable or useful information, but it must also relay information of at least equal quality. In other words, an influential account must act both as an authority and as a hub, strongly hinting that online influence is of inherently higher-order nature.

Similarly, the use of hashtags by μ^I and μ^M is important because of their semantic value. Moreover, it can be argued that an account which strategically or creatively places hashtags is likely to attract attention and probably followers, possibly explaining why μ^M is close to μ^S or μ^A and revealing some correlation between hashtag variability and similarity patterns between Φ_k and Ψ_k. If true, this would add to the evidence suggesting that functional and structural features should be carefully merged to hybrid influence metrics.

5 Experimental evaluation

5.1 Execution time requirements

Besides the correctness and the expressive power of the proposed analytics, an efficient implementation is also necessary to demonstrate their potential. Two scenaria were examined following the procedures specified in Drakopoulos et al. (2016a). According to the first scenario (s1), our application, composed of the database and the analytics, was the only one running and, thus, had unlimited access to the resources shown in Table 11. This led to the establishment of the baseline performance. Then, the system was running with an average workload of 0.5 (s2), a moderate value, and our application was run 11 times in order for the information shown in Table 12 to be collected, where the measurements of the first execution were ignored. These figures relate to analytics only as the social graph was already collected at an earlier date. Note that the values of the analytics themselves were the same in each run regardless of the total execution time and system workload, as the metrics are deterministic by construction.

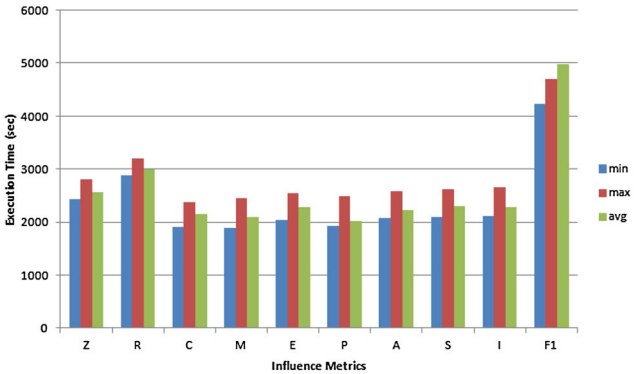
Table 11 System specifications

Property	Value
CPU	AMD Athlon X4@4 GHz
CPU cores	4
Hard disk	1 TB
L2 cache	4 GB
Memory	16 GB
OS	Ubuntu 16.04
Swap partition size	16 GB

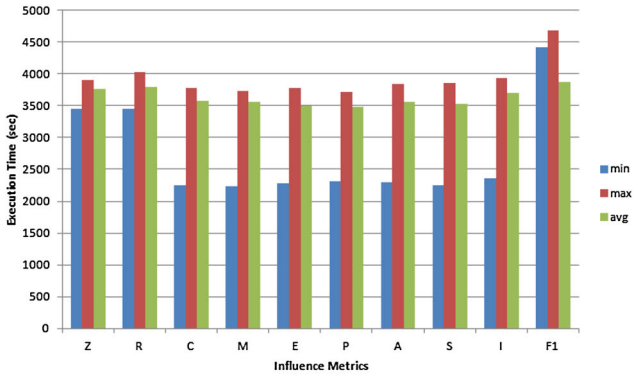
In general, the computation of baseline metrics μ^Z and μ^R as well as μ^{F1} was slower. This can respectively be attributed to the fact that μ^Z and μ^R entail the solution of a large and sparse linear system, whereas μ^{F1} requires the computation of two metrics. On the contrary, the remaining metrics rely on matrix-vector multiplications $\mathbf{Ax}$ which is potentially an order of magnitude quicker than a linear system solution $\mathbf{A}^{-1}\mathbf{x}$. Also, the standard deviations in each scenario were very similar with those in the second scenario being consistently higher and distinctive than those in the first. Finally, in the first scenario the mean value was approximately equidistant to the lower and the higher times, whereas in the second the mean is closer to the maximum value. The information of Table 12 is repeated in Fig. 2a, b for clarity.

Table 12 Execution time for influence rankings (**m**in, ma**x**, **a**vg, **s**td in sec)

s1	μ^Z	μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
m	2441	2882	1916	1893	2045	1933	2081	2103	2112	4234
x	2816	3194	2378	2450	2542	2481	2589	2623	2653	4692
a	2564	3004	2157	2093	2285	2017	2234	2302	2285	4982
s	111	113	111	111	112	110	113	112	112	42
s2	μ^Z	μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
m	3447	3459	2246	2231	2284	2315	2298	2256	2352	4418
x	3911	4025	3773	3725	3777	3721	3844	3856	3929	4678
a	3760	3791	3572	3568	3501	3489	3563	3524	3699	3871
s	117	117	118	117	116	116	116	116	118	39



(a) Time (s1).



(b) Time (s2).

Fig. 2 Total execution time for each influence ranking scheme

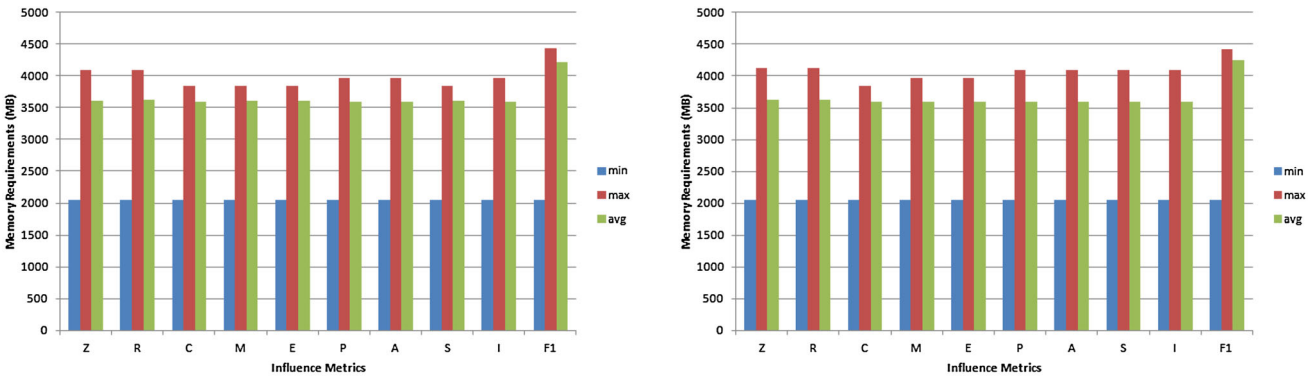
Table 13 Memory requirements for influence rankings (**m**in, ma**x**, **a**vg, **s**td in MB)

s1	μ^Z	μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
m	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048
x	4096	4096	3840	3840	3840	3968	3968	3840	3968	4432
a	3611	3617	3592	3602	3604	3589	3591	3606	3596	4212
s	112	111	112	112	112	111	113	113	113	110
s2	μ^Z	μ^R	μ^C	μ^M	μ^E	μ^P	μ^A	μ^S	μ^I	μ^{F1}
m	2048	2048	2048	2048	2048	2048	2048	2048	2048	2048
x	4124	4124	3840	3968	3968	4096	4096	4096	4096	4416
a	3619	3622	3597	3596	3601	3596	3601	3593	3597	4254
s	113	112	111	112	112	111	112	111	112	110

5.2 Memory requirements

The *top* command was configured in batch mode. Memory use was updated every 10 seconds, a resolution smoothing out any spikes or outliers. The measurments are summarized in Table 13 and in Fig. 3a, b.

The two baseline metrics μ^R and μ^Z had consistently more memory requirements as denoted by the high average memory use and the simultaneous low standard deviation. On the contrary, the remaining seven metrics were less memory intensive in general with occasional spikes,



(a) Memory (s1). **(b)** Memory (s2).

Fig. 3 Memory requirements for each influence ranking scheme

probably while caching columns of the adjacency matrix $\mathbf{M}$. μ^P, μ^S, and μ^A are more demanding, perhaps because they need to compute the cardinality of large sets of followers and followed. At a relative distance comes μ^I followed closely by μ^C and μ^E. Their common denominator is that they only rely on a large set and thus, place less strain memorywise. At the last place comes μ^M, probably as a result of dealing only with the smaller set of hashtags.

In contrast to the time measurements, memory requirements, including the standard deviations, are approximately the same in both scenaria. This can be attributed to the fact that moving memory blocks back to main memory or even to the swap partition may be time consuming but imposes no additional memory constraints. Therefore, the memory allocated to our application depends solely on the data to be processed as well as on any extra memory required by the analytics themselves. As these memory factors remain constant across executions, the total memory needs to remain constant as well.

The above findings depend heavily on the various dataset features. For instance, in a network with lower sets of followers but with more tweets, retweets, and hashtags, a commonplace characteristic in news or professional communication networks (Russell 2013), memory usage patterns might be different.

6 Conclusions and future work

Ranking influential Twitter accounts and evaluating the quality of influence metrics are addressed in this article. A general algebraic scheme for deriving higher-order Twitter influence rankings from first-order rankings was proposed. At its heart lies a convex combination of the score of a given account and those of its followers. Moreover, the analytical framework of Drakopoulos et al. (2016b) for assessing the performance of an influence metric was extended with probabilistic tools from various field including information theory, data mining, and psychometrics. The framework was implemented in the Python ecosystem.

In order to reach meaningful conclusions regarding the performance analysis of Twitter influence rankings the Katz metric μ^Z and the TunkRank μ^R were used as baselines. The former is representative of structural rankings, whereas the latter is a functional one. Analysis indicates that among the seven higher-order metrics derived by the proposed scheme, namely μ^C, μ^M, μ^E, μ^P, μ^A, μ^S, μ^{F1}, and μ^I, μ^I performs well compared to both μ^Z and μ^R, reflecting the fact that it combines structural and functional features, although it is among the more demanding metrics. Another interesting finding was that μ^E, μ^M, and μ^C had similar behavior to μ^R. The fact that μ^I outperforms the other digital influence rankings can be attributed to the diverse factors it combines, which capture a major part of Twitter activity.

Possible algorithmic research can be conducted toward developing sophisticated Twitter hybrid metrics. Moreover, scalability is an issue that should be taken into consideration. Also, the proposed metrics should be applied to networks from other domains or to networks from multiple domains, as for instance to a combination of educational and news accounts. Concerning the implementation of the proposed rankings, any sparsity patterns of the adjacency matrix should be exploited in ranking computation. Matrix-free methods or parallel matrix-vector multiplication, perhaps in combination with advanced indexing may be a way to achieve lower execution times.

Moving beyond the proposed analytics, Twitter influence metrics should integrate reputation and trustworthiness in social media as well as their evolution over time as a means to curtail online trolling and the spreading of fake news, such as topics emphasized during the 2016 US elections. Finally, live analytics will shed more light to the actual Twitter structure and will be used to predict online events in real time.

References

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the 4th ACM WSDM. ACM, pp 65–74

Benzi M, Boito P (2010) Quadrature rule-based bounds for functions of adjacency matrices. Linear Algebra Appl 433(3):637–652

Benzi M, Klymko C (2015) On the limiting behavior of parameter-dependent network centrality measures. SIAM J Matrix Anal Appl 36(2):686–706

Bertot JC, Jaeger PT, Grimes JM (2010) Using ICTs to create a culture of transparency: e-government and social media as openness and anti-corruption tools for societies. Gov Inf Q 27(3):264–271

Bi Z, Faloutsos C, Korn F (2001) The DGX distribution for mining massive, skewed data. In: Proceedings of the seventh ACM SIGKDD. ACM, pp 17–26

Bickmore T, Cassell J (2001) Relational agents: a model and implementation of building user trust. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 396–403

Blackmore S (2000) The meme machine. Oxford University Press, Oxford

Bodnar T, Tucker C, Hopkinson K, Bilén SG (2014) Increasing the veracity of event detection on social media networks through user trust modeling. In: 2014 IEEE international conference on big data. IEEE, pp 636–643

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92:1170–1182

Bouguessa M, Dumoulin B, Wang S (2008) Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, KDD '08, pp 866–874

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 media

Dawkins R (2006) The selfish gene, 3rd edn. Oxford University Press, Oxford

Drakopoulos G (2016) Tensor fusion of social structural and functional analytics over Neo4j. In: Proceedings of the 6th international conference of information, intelligence, systems, and applications. IEEE, IISA 2016

Drakopoulos G, Kanavos A (2016) Tensor-based document retrieval over Neo4j with an application to PubMed mining. In: Proceedings of the 6th international conference of information, intelligence, systems, and applications. IEEE, IISA 2016

Drakopoulos G, Megalooikonomou V (2016) Regularizing large biosignals with finite differences. In: Proceedings of the 6th international conference of information, intelligence, systems, and applications. IEEE, IISA 2016

Drakopoulos G, Baroutiadi A, Megalooikonomou V (2015) Higher order graph centrality measures for Neo4j. In: Proceedings of the 6th international conference of information, intelligence, systems, and applications. IEEE, IISA 2015

Drakopoulos G, Kanavos A, Makris C, Megalooikonomou V (2016a) Finding fuzzy communities in Neo4j. In: Howlett RJ, Jain LC (eds) Smart innovation, systems, and technologies. Springer, Berlin

Drakopoulos G, Kanavos A, Tsakalidis A (2016b) Evaluating Twitter influence ranking with system theory. In: Proceedings of the 12th international conference on web information systems and technologies, WEBIST 2016

Estrada E, Higham DJ (2010) Network properties revealed through matrix functions. SIAM Rev 52(4):696–714

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslov Math J 23(2):298–305

Gao Y (2005) Factors influencing user trust in online games. Electron Libr 23(5):533–538

Golbeck J (2009) Trust and nuanced profile similarity in online social networks. ACM Trans Web 3(4):12

Kafeza E, Kanavos A, Makris C, Chiu D (2013) Identifying personality-based communities in social networks. In: Legal and social aspects in web modeling (Keynote Speech) in conjunction with the international conference on conceptual modeling (ER), LSAWM

Kafeza E, Kanavos A, Makris C, Vikatos P (2014) T-PICE: Twitter personality-based influential communities extraction system. In: IEEE International Congress on Big Data, pp 212–219

Kanavos A, Perikos I, Vikatos P, Hatzilygeroudis I, Makris C, Tsakalidis A (2014a) Conversation emotional modeling in social networks. In: 26th IEEE international conference on tools with artificial intelligence. ICTAI, pp 478–484

Kanavos A, Perikos I, Vikatos P, Hatzilygeroudis I, Makris C, Tsakalidis A (2014b) Modeling retweet diffusion using emotional content. In: Artificial intelligence applications and innovations. AIAI, pp 101–110

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Leskovec J (2011) Social media analytics: tracking, modeling and predicting the flow of information through networks. In: Proceedings of WWW 2011. ACM, pp 277–278

Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1361–1370

Leskovec J, Rajamaran A, Ullman JD (2014) Mining of massive datasets, 2nd edn. Cambridge University Press, Cambridge

Li J, Shiu WC, Chang A (2009) On the Laplacian Estrada index of a graph. Appl Anal Discr Math 3:147–156

Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I et al (2011) The Arab Spring-the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5:31

Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, Czajkowski G (2010) Pregel: a system for large-scale graph processing. In: Proceedings of ICDM 2010. ACM, pp 135–146

Manicas PT (1991) History and philosophy of social science. PhilPapers

Mehta R, Mehta D, Chheda D, Shah C, Chawan PM (2012) Sentiment analysis and influence tracking using twitter. Int J Adv Res Comput Sci Electron Eng 1(2):73–79

Newman ME (2005) A measure of betweenness centrality based on random walks. Soc Netw 27(1):39–54

Okamoto K, Chen W, Li XY (2008) Ranking of closeness centrality for large-scale social networks. In: International workshop on frontiers in algorithmics. Springer, pp 186–195

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the Wweb

Pal A, Counts S (2011) Identifying topical authorities in microblogs. In: Proceedings of the Fourth ACM international conference on web search and data mining. ACM, WSDM '11, pp 45–54

Panzarino O (2014) Learning cypher. PACKT publishing, Birmingham

Razis G, Anagnostopoulos I (2014) InfluenceTracker: rating the impact of a twitter account. In: Proceedings of AIAI, pp 184–195

Rivest RL, Vuillemin J (1976) On recognizing graph properties from adjacency matrices. Theor Comput Sci 3(3):371–384

Robinson I, Webber J, Eifrem E (2013) Graph databases. O'Reilly, Sebastopol

Rogers EM, Beal GM (1957) The importance of personal influence in the adoption of technological change. Soc F 36:329

Russell MA (2013) Mining the social web: analyzing data from Facebook, Twitter, LinkedIn, and other social media sites, 2nd edn. O'Reilly, Sebastopol

Smith AN, Fischer E, Yongjian C (2012) How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? J Interact Mark 26(2):102–113

Tong H, Prakash BA, Eliassi-Rad T, Faloutsos M, Faloutsos C (2012) Gelling and melting large graphs by edge manipulation. In: Proceedings of the 21st CIKM. ACM, pp 245–254

TunkRank (2015) http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank

Turner JC (1991) Social influence. Thomson Brooks/Cole Publishing Co, Pacific Grove

Weng J, Lim EP, Jiang J, He Q (2010) TwitterRank: Finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM international conference on web search and data mining. ACM, pp 261–270

Zamparas V, Kanavos A, Makris C (2015) Real time analytics for measuring user influence on twitter. In: Proceedings on the 27th international conference on tools with artificial intelligence. IEEE, pp 591–597

Zimbardo PG, Leippe MR (1991) The psychology of attitude change and social influence. McGraw-Hill Book Company, New York