Procedia Computer Science

Volume 51, 2015, Pages 2435–2444

ICCS 2015 International Conference On Computational Science

Quantifying complementarity among strategies for influencers' detection on Twitter*

Alan Neves, Ramon Vieira, Fernando Mourão, and Leonardo Rocha

Universidade Federal de São João del Rei, São João del Rei, Minas Gerais, Brasil
{aneves,ramonv,fhmourao,lcrocha}@ufsj.edu.br

Abstract

The so-called influencer, a person with the ability to persuade people, have important role on the information diffusion in social media environments. Indeed, influencers might dictate word-of-mouth and peer recommendation, impacting tasks such as recommendation, advertising, brand evaluation, among others. Thus, a growing number of works aim to identify influencers by exploiting distinct information. Deciding about the best strategy for each domain, however, is a complex task due to the lack of consensus among these works. This paper presents a quantitative study of analysis among some of the main strategies for identifying influencers, aiming to help researchers on this decision. Besides determining semantic classes of strategies, based on the characteristics they exploit, we obtained through PCA an effective meta-learning process to combine linearly distinct strategies. As main implications, we highlight a better understanding about the selected strategies and a novel manner to alleviate the difficulty on deciding which strategy researchers would adopt.

Keywords: Social Media Environments, Information Diffusion, Influencer Detection

1 Introduction

The widespread use of social media applications, such as blogs, microbloggings and social networks, has allowed an increasing among of people to diffuse on the Web thoughts, opinions, discussions and reviews about distinct topics [4]. Hence, researchers on information diffusion possess nowadays a rich and huge amount of data to understand better this diffusion process. Indeed, understanding the diffusion of information has implications for several tasks, such as recommendation, advertising, opinion pools, brand evaluation, among others [1, 23]. Efforts on this direction, frequently, point out the so-called influencers as a main character on this diffusion process. An influencer is a person with the ability to persuade people, affecting their actions and behavior [17]. Besides their sociological role, determining which information will be spread, as well as the diffusion speed and reach, influencers have an important economic

*This work was partially supported by CNPq, CAPES, Fapemig, and INWEB.

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.428

role. Recent studies have demonstrated that word-of-mouth and peer recommendation, which might be dictated by influencers, are effective forms of advertising [1]. Thus, a growing number of studies have focused on identifying influencers on distinct social media environments.

Despite the identification of influencers has been an active research area, it has proven to be a challenging task [4]. We highlight three main difficulties related to this task. First, it is not clear which characteristics are relevant to determine a person as being an influencer. While some studies exploit social structure of networks, other focus on patterns of propagation in the network or basic statistics of each person, such as number of posts spread over his/her contacts, total of answered messages, among others. The second difficulty is how to determine which characteristics must be considered in order to determine an influencer, according to a specific goal (i.e. economic, social, etc.) and the available information (connection between users, information route, etc.). Third, a major difficulty is how to assess the effectiveness of the proposed methods in identifying all influencers of each domain. Since there is no ground truth about the actual influencers in many real domains, quality assessment became unclear and controversial in the literature [1, 4]. Distinct proposals have emerged over time but no consensus on how to address this task was reached. Further, to the best of our knowledge, there is no work in the literature that analyzes and compares the main strategies for identifying influencers.

This work performs a quantitative study of analysis and comparison among some of the main strategies for identifying influencers in the literature. We intend to help researchers on information diffusion to understand better the effect of distinct strategies in each domain and make a better decision about which of these strategies to use. In this sense, first, we surveyed several efforts on the literature and proposed a new taxonomy for these efforts, based on common premises and characteristics they exploit. Specifically, we evaluated six highly referenced strategies: PageRank, PCC, ProfileRank, Effective Readers, and two basic statistical summaries. Thus, we identified three relevant classes. The first class comprises strategies exploiting network structures, which is represented by *PageRank*$^{\text{TM}}$ algorithm [15] and PCC [9]. PageRank calculates a score of influence for each vertex in a directed graph using only relationships and propagation on the network. PCC proposes a centrality-based metric to determine influential neighborhoods in a network. Our second class refers to strategies focused on content and flow, which is represented by the strategies ProfileRank [18] and Effective Readers [11]. While ProfileRank models information diffusion just considering the temporal order in which the messages are propagate on a social network, Effective Readers evaluates information diffusion as a cascade effect that topics have among users. The third class consists of strategies that exploit statistical summary of user actions and attributes. The strategies belonging to this class are Number of social ties [23] and Number of propagated posts [1].

In order to evaluate the proposed taxonomy, we measure the level of agreement among the selected strategies. We conducted all analyses on data samples from Twitter, given its relevance for information diffusion on the Web. We applied each strategy S_a on these data sample and derived a descending ordered list of Top-50 influencers, according to the scores defined by S_a. Then, we used the generalized version Kendall's Tau metric [6] as measurement of agreement between pair of lists. Our analyses confirmed that lists derived from strategies belonging to the same class present the highest level of agreement. Since these analyses demonstrated high variance on the scores derived by each strategy, we also used Principal Component Analysis (PCA) for extracting useful and orthogonal information modeled by them [16]. Through PCA, first, we analyzed the complementarity of information modeled by strategies belonging to the same class. We found that strategies belonging to the same class are strongly correlated, once again, corroborating the proposed taxonomy. Then, we evaluated the complementarity among strategies belonging to distinct classes. Our interest was to determine how redundant strategies

derived from distinct fields and theories are. In this case, first we observe the resulting lists of influencers, generated by each strategy, are very different from each other. Furthermore, we may interpret the use of PCA as a meta-learning strategy that combines linearly distinct strategies in order to compose a single one carrying the useful information to identify influencers. Moreover, the use of PCA alleviates the difficulty on deciding which strategy researchers would adopt to get a proper list of influencers. As main contributions, we highlight:

1. *Survey and organization of efforts in the literature into semantic class (i.e., a new taxonomy), based on assumptions and characteristics exploited*;

2. *Analysis of intra-class complementarity using PCA, which showed high level of agreement among strategies of the same class*;

3. *Analysis of inter-class complementarity, which showed that PCA can be used as a meta-learning process combining linearly the strategies that consider different characteristics.*

2 Related Work

There are in the literature several works that aim to identify influencers in social media environments [1, 23]. It is noteworthy the diversity of domains and fields related to these works, ranging from sociological to biological researches [17]. Considering the domain of study, we grouped most of the efforts into three main groups.

First, we found many works focused on sociological theories. In [19], the authors proposed a method based on the notion of social capital, which measures the ability of bonding (i.e., similar) and bridging (i.g., diverse) people to cooperate and communicate in a network. In turn, [9] proposed a centrality measure, named Principal Component Centrality (PCC), to determine influential neighborhoods in a network. In [10], the authors showed that the most efficient influencers are those located within the core of the network, identified by a k-core decomposition analysis. Also, [21] reviewed 10 techniques used to identify influencers, presenting advantages/disadvantages on each technique.

The second group is related to mathematical/computational modeling of the problem. Studies on this group range from simple statistical analysis of activity logs to graph-based strategies [1, 23, 9]. Usually, there is a high intersection between the type of information exploited by strategies from the first group and those exploited in the second one. However, strategies from the second group tend to combine simple information of sociological structures with additional information, such as content and time. Indeed, some works argue that the user's actual influence is related to his/her relationships and the temporal order of the published content [11, 18]. In this sense, [18] takes into account the temporal order of message diffusion, without considering the relationships for deriving a user's influence score. In turn, [15] adapted the PageRank algorithm to derive an influence score for each user, modeled as a node in a directed graph. Other works model the problem of finding influencers as an influence maximization problem, where the goal is to find the top-k nodes such that the average infection spread is maximized [12].

Finally, we point out as the third relevant group works based on economic theories [20, 7, 13]. Theories in this group are, usually, focused on viral marketing and brand adoption. The goal is to understand the mechanisms that lead to large-scale chain-reaction of influence, with a very small marketing cost. Using a threshold model, [20] studied the opinion leadership on the diffusion of innovations. In [22], the aim was to clarify insights on new product diffusion in a population of influencers, through mathematical formalism. Further, [13] used a neural artificial network to find influential reviewers for word-of-mouth marketing. Also, [7] proposed the *Law of the Few* to explain the role of influencers in social groups, which states that a majority of individuals get most of their information from a very small subset of users, the influencers.

Since our case studies are data samples from Twitter, we also reviewed some main efforts for identifying influencers on it. Indeed, Twitter has been the focus of several researches on information diffusion [11]. Many of the aforementioned works use this microblogging system as case study, such as [4, 23, 1]. Also, Twitter's singularities motivated the proposal of novel strategies specifically for it. [3] adapted the k-shell decomposition algorithm to exploit followers relationship. [1] conducted an empirical analysis on the attributes and roles of influencers.

A main drawback related to the numerous works existing in the literature, however, is that there is no consensus among them [23]. Determining which strategy might provide a proper solution in each domain is not trivial, since quality evaluation is complex and unclear [4].Further, to the best of our knowledge, there is no work focused on quantifying the similarity among distinct strategies. Thus, this work aims to analyze and compare quantitatively some of the main strategies in the literature, instead of proposing new ones. Moreover, by applying PCA to determine not correlated information, we have a simple way to combine these strategies.

3 A new taxonomy for influencer detection strategies

In this section, we present a new taxonomy for efforts on identifying influencers, based on the distinct assumptions and characteristics exploited in the literature. Thus, we aim to determine semantic classes of strategies that exploit the same type of characteristics. Such classes would help us in our goal of analyzing and comparing distinct strategies.

In order to define these classes, we propose a simple and non-automatic methodology of classification. First, we surveyed the main related works published or referenced in the past few years. Second, we extracted from these works the main characteristics they exploit in order to determine influencers. Third, we grouped these characteristics according to the type of information modeled. Finally, we classified the works according to the groups defined in the last step. Note that, this methodology can be applied to any set of strategy existing in the literature. Further, the resulting groups would depend on the set of evaluated strategies. Also, we highlight that we do not intend to provide a closed set of classes in this work. Instead, we propose a way to extend the taxonomy as new strategies appear.

3.1 Evaluated Strategies

In this section, we describe in detail the six highly referenced strategies we selected.

- **PCC:** Principal Component Centrality (PCC) [9] is a strategy of influencer detection based on centrality, a measurement of relevance for users in a network that takes into account the neighborhood of each user. PCC extends the **E**ingen**V**alue **C**entrality (EVC) metric [2], by exploiting the k dominant characteristics in a graph. These characteristics are determined through a Singular Value Decomposition (SVD) of the adjacent matrix representing the social network [8]. Based on this decomposition and on the Hadamard operator ($\odot$) [5], the PCC's value is derived according to Equation 1. In this formula, k is the number of characteristics (i.e., eigenvectors) to be exploited and U and S are two matrices obtained through SVD. By applying PCC on an adjacent matrix, we derive a centrality measure for each user. The higher this centrality, the higher the user's influence.

$$C_p = \sqrt{(U_{n\times k} \odot U_{n\times k})(S_{k\times 1} \odot S_{k\times 1})} \quad (1)$$

- **PageRank:** PageRank is a well-know algorithm proposed in [15] and commercially exploited by Google^{TM} to determine global ranks for Webpages. Basically, it calculates the propagation of influence among nodes on a directed graph. In order to identify influencers, nodes on this graph represent users and edges their social relationships. Thus, the PageRank value (PR) of each user u_i is given by Equation 2, where M_i is the set of users connected to user u_i; $L(j)$ is the size of u_j's neighborhood; α is the damping factor (with default value 0.85); and N is the number of users in the graph. The PageRank value represents the probability of reaching each user in the network. The higher this probability, the higher the user's influence.

$$PR(u_i) = \alpha \sum_{j \in M_i} \frac{PR(u_j)}{L(j)} + \frac{1-\alpha}{N} \tag{2}$$

- **Effective Readers:** This strategy considers relationship among users and the temporal order in which information reach each user. The authors observed that information spread faster at the initial moments of its diffusion process. Based on this observation, the authors proposed the Effective Readers score, which is the number of distinct users who received an information for the very first time from a specific user. Assuming that users read their messages chronological as the latter arrive, given a message m published by a user u_i, the Effective Readers $ER_0(m, u_i)$ is defined by Equation 3, where F is a binary flag of status for each user u_j. F is initialized to zero and it is set to 1 at the first moment u_j gets m from his/her neighborhood. The influence score IF_0 assigned to u_i is defined as the sum of ER_0 values derived for the set $T(u_i)$ of all messages published by u_i, such as show by Equation 4.

$$ER_0(m, u_i) = \sum_{u_j \in follower(u_i)} \overline{F(u_j, m)} \tag{3}$$

$$IF_0(u_i) = \sum_{m \in T(u_i)} \parallel ER_0(m, u_i) \parallel \tag{4}$$

- **ProfileRank:** This strategy exploits a cyclic definition of relevance and influence [18]. Influential users spread relevant content and relevant content are disseminated by influential users. ProfileRank sort both users and content according to scores of influence and relevance. It models information diffusion through the relationship among users and content over time by using two matrices M and L. M is a matrix with dimensions $|U| \times |C|$ that represents the content set C created by the user set U. ProfileRank initializes each position $M_{i,j}$ as $\frac{1}{q_i}$, where q_i is the number of pieces of contents u_i has created or propagated. In turn, matrix L has dimensions $|C| \times |U|$, where $L_{i,j} = 1$ if the user u_j created the piece of content c_i and $L_{i,j} = 0$ otherwise. Based on these matrices, scores of content relevance (r) and user influence (p) are defined according to Equations 5 and 6, respectively. Again, d stands for the damping factor and, in this case, it is used to prevent strongly connected subgraphs. I is the identity matrix and V is a uniform vector.

$$r = (1-d)V(I - dLM)^{-1} \tag{5}$$

$$p = (1-d)V(I - dML)^{-1} \tag{6}$$

- **Number of Followers:** one of the simplest and most frequently used strategies designed specifically for Twitter [4]. A follower is a user that want to get updated about actions from the user he/she is following. Basically, some works count the number of followers each user has in the system. The higher this number, the higher the user influence on the network.

- **Number of Retweets:** another simple and very referenced strategy designed for Twitter [23]. It summarizes the number of tweets published by a user that were forwarded by other users (i.e., retweets). The higher the number of retweets related to a given user, the higher this user's influence on the network.

3.1.1 Resulting Taxonomy

Following the methodology presented, the third step corresponds to group the characteristics according to the type of information modeled, and, as fourth step, the strategies are classified according to the groups previous defined. Based on the subset of evaluated strategies, we identified three main classes, as follows.

The first class comprises strategies that take into account Network Structures (NS) only. We found in this group the *PageRank*^{TM} algorithm [15], which calculates a score of influence for each vertex in a directed graph using only relationships and propagation on the network. PCC [9] is the second strategy in this class. Again, it proposes a centrality-based metric to determine influential neighborhoods in a network.

Strategies belonging to the second class exploit Content and Flow (C&F) to determine influencers. We observe in this class ProfileRank [18] and Effective Readers [11]. While ProfileRank models information diffusion just considering the temporal order in which the messages are propagate on a social network, Effective Readers evaluates information diffusion as a cascade effect that topics have among users.

Finally, the third class corresponds to strategies focused on Statistical Summary (SS) of activity logs of users. Basically, they are strategies that aim to determine the influential level of users by just summarizing some attributes of users, such as the Number of Followers (#F) and Number of Retweets (#R).

3.2 Taxonomy Evaluation

In order to evaluate the proposed taxonomy, we conducted an experiment on which we measure the level of agreement among the selected strategies. We conducted all analyses on data samples from Twitter, given its relevance for information diffusion on the Web. In the following subsections, we describe in detail these data samples, the experiments and the results achieved.

3.2.1 Data Samples

We collected two distinct samples from Twitter. The first one (Dataset1) refers to tweets coming that the Brazilian city Belo Horizonte and spans the period from 10/30/2014 to 11/06/2014. During this period the two biggest soccer team from Belo Horizonte were playing the finals of the two biggest Brazilian soccer championship (i.e., Copa do Brasil and Campeonato Brasileiro). The second data sample (Dataset2) comprises tweets coming from the biggest Brazilian city, São Paulo, published from 12/24/2014 to 01/05/2015 that corresponds to the period of Christmas and New Years Eve celebrations. In both collections, we collected tweets related to the top 10 topics most discussed (i.e., the *TopTrends*) in the previous 10 minutes of gathering. We also collected public information about all users related to a tweet, such as the tweet's writer, his/her followers and friends. We used the *APIStreaming*[1] to gather the tweets and the API REST to gather all user's public information[2]. Table 1 details both data samples.

Table 1: Summary of the data samples used in all experiments.

Collection	# users	# tweets
Dataset 1	1,253	3,248
Dataset 2	3,800	1,302

[1]https://dev.twitter.com/docs/streaming-apis
[2]https://dev.twitter.com/docs/api

3.2.2 Experimental evaluation

To determine the level of agreement among distinct strategies, we conducted a pairwise comparison among results generated by each strategy. In this sense, we applied each strategy S_a on our data samples and derived a descending ordered list of Top-50 influencers, according to the scores defined by S_a. Then, we used the generalized version of Kendall's Tau metric [6] with penalty parameter $p = 0$ as measurement of agreement between pair of lists. The higher the Kendall's Tau between two lists, the more they agree. To ease comparisons, we normalized the results obtained, such as suggested by [14]. Table 2 shows the results of this analysis.

Table 2: Pairwise comparison of lists of Top-50 influencers generated by strategy of influencer identification, using Kendall's Tau metric.

	PCC	PageR	ER	PR	#R	#S
PCC	1.00	0.77	0.24	0.09	0.32	0.29
PageR	0.78	1.00	0.29	0.09	0.34	0.32
ER	0.23	0.28	1.00	0.35	0.40	0.68
PR	0.09	0.09	0.36	1.00	0.05	0.08
#R	0.32	0.34	0.41	0.05	1.00	0.62
#S	0.29	0.32	0.68	0.08	0.62	1.00

A - Dataset 1

	PCC	PageR	ER	PR	#R	#S
PCC	1.00	0.72	0.36	0.15	0.02	0.22
PageR	0.68	1.00	0.31	0.10	0.23	0.34
ER	0.36	0.32	1.00	0.72	0.01	0.13
PR	0.15	0.11	0.73	1.00	0.00	0.02
#R	0.02	0.23	0.01	0.00	1.00	0.43
#S	0.21	0.32	0.13	0.02	0.43	1.00

B - Dataset 2

As highlighted in the tables, our analysis confirmed that lists derived from strategies belonging to the same class present the highest level of agreement on both datasets, which correspond to our first contribution on this paper. An exception refers to the Effective Readers (from class C&F) on Dataset1, which exhibited higher levels of agreement with the strategy Number of Followers (from class SS). This behavior is related to the collection size and limitations inherent to these strategies. Since Dataset1 spans a short period of time, there are few users in this collection with high number of relationships (e.g., higher than 15). In this kind of scenario, the Effective Readers behaves similarly to the Number of followers because there are not enough connections in the network to model the information propagation.

Based on the these results, a relevant question arises: is it possible to derive a single metric from each class that combines useful and distinct information captured by each strategy of the same class? Next, we present the experiments performed in order to answer this question.

4 Analysis of intra-class complementarity

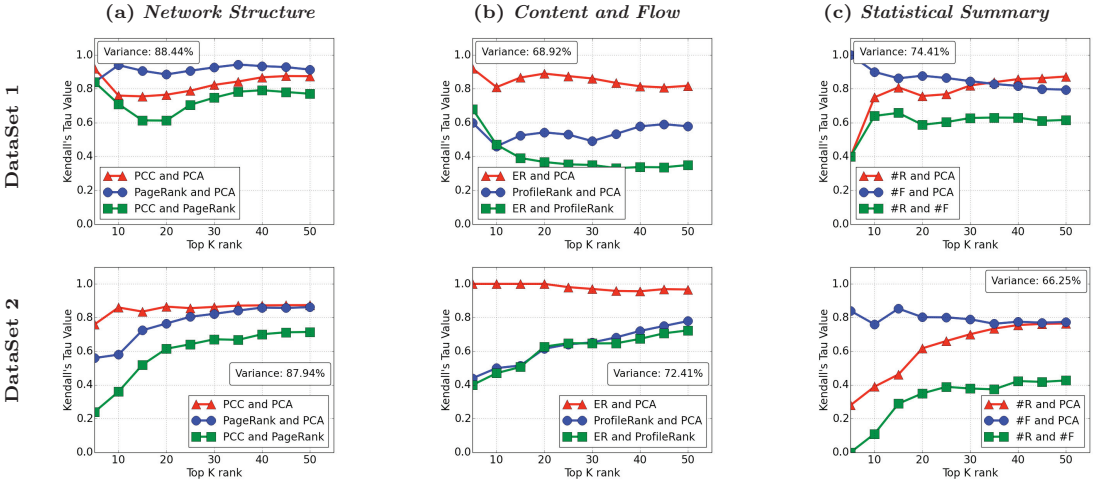
In this section, we analyze the complementarity of information modeled by strategies belonging to the same class. The experiments conducted in Section 3.2.2 demonstrated high variances on the scores derived by each strategy. This observation motivated us to use Principal Component Analysis (PCA)[16] for extracting useful and orthogonal information modeled by them. Thus, instead of deciding among strategies from the same class, researchers may use the combination of these strategies performed by PCA.

PCA is a multivariate statistical technique that exploits the variability structure of the data [16]. Its main idea consists of reducing the dimensionality of a dataset that exhibit a large number of correlated variables, while capturing as much as possible the original data variability. PCA performs this reduction by transforming the variables into a new set of orthonormal variables, named Principal Components. These components are decreasingly ordered by the amount of the variability each one models. In our analyses, we run PCA using an input matrix P. P has dimensions $|U| \times |S|$, where U is the user set and S is the set of evaluated strategies. Each position $P(u_i, s_j)$ refers to the influence score that strategy s_j assigned to user u_i. Since one of our goals is to derive a single metric that captures all non-redundant information modeled by each class, we evaluated only the first principal component of PCA. The first component

might be interpreted as a liner combination of the strategies. Thus, we derived a new influence score for each user u_i that is a linear combination of all scores given to u_i by distinct strategies of each class. As our taxonomy has three class, we created three distinct input matrices P, each one containing only the strategies of each class (i.e., $|S| = 2$).

Again, for each class, we derived a descending ordered list of Top-K influencers, according to the new scores defined by PCA and compared this list with the ones related to each strategy belong to the same class, using the Kendall's Tau metric. The goal is to measure the agreement between the new list generated by PCA with the others strategies of the same class. Figure 1 presents the results of this analysis when varying the Top-K influencers from 5 to 50.

Figure 1: Analysis of complementarity of information modeled by distinct strategies belonging to the same class using PCA.



Observe that, in most of the cases, strategies belonging to the same class (i.e., green lines) present levels of agreement higher than 40%. Additionally, when comparing the Top-K influencers derived from PCA with other strategies belonging to the same class the levels of agreement are even higher in both collections. For all datasets and classes, we found at least 80% of agreement between PCA and one of the strategies. These observations have two main implications. First, strategies belonging to the same class are strongly correlated, once again, corroborating the proposed taxonomy. Second, the Top-K influencers identified by PCA were able to synthesize properly the characteristics exploited by each class. The first principal component was able to capture more than 66% of the data variability in all cases (i.e. the Variance presented in graphics). Furthermore, we may interpret the use of PCA as a meta-learning strategy that combines linearly distinct strategies of a well defined class in order to compose a single one carrying the useful information to identify influencers. Indeed, this is one of the main contributions of this work.

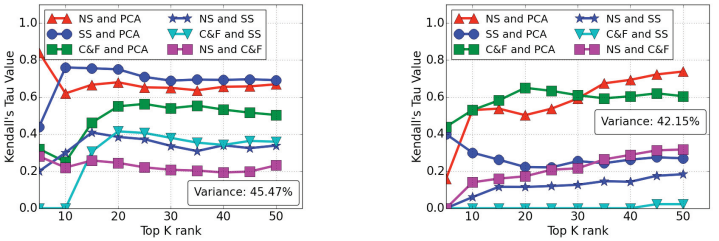
5 Analysis of inter-class complementarity

An important question raised by the foregoing discussion refers to the complementarity among strategies belonging to distinct classes and how PCA would be able to combine strategies that exploit characteristics from distinct nature. In this section, we evaluate this question.

Such as performed in Section 3, we created an input matrix P of users per strategies for each data sample. The position $P(u_i, s_j)$ represents the influence score assigned by strategy s_j

to user u_i. We used all evaluated strategies to compose P (i.e., $|S| = 6$). Then, we applied PCA on P deriving a single influence score per each distinct user, which represents a combination of the scores originally defined by each strategy. For each data sample, we derived a descending ordered list of Top-50 influencers, according to the new score defined by PCA. We contrasted the list generated by PCA over all strategies against the lists resulting from PCA applied over each class individually. Using the Kendall's Tau metric for determining levels of agreement, we varied the Top-K influencers from 5 to 50. The results of this analysis are presented in Figure 2.

Figure 2: Analysis of complementarity of information modeled by strategies belonging to different classes using PCA



(a) Dataset 1. (b) Dataset 2.

We observe that lists derived by PCA on distinct classes have low agreement among them (i.e., smaller than 40%). This result demonstrates that as strategies belonging to distinct classes exploit different information, the resulting Top-K influencers of each strategy differ from each other, depicting the difficulty in selecting the best strategy to use due to lack of consensus. On the other side, the list generated by PCA applied to all strategies, in most of the cases, synthesized better the information modeled by the different classes, presenting levels of agreement higher than 50%. The variance captured by the first principal component tends to be smaller (around 40%) given the larger number of non-aligned strategies. Thus, PCA may be also used as a meta-learning strategy to combine linearly strategies with distinct nature. We believe this is the main contribution of this work, since PCA may alleviate he difficulty on deciding which strategy researchers on information diffusion would adopt to get a proper list of influencers.

6 Conclusions and Future Works

In this paper, we presented a quantitative study of analysis and comparison among some of the main strategies for identifying influencers on social media applications [9, 18, 11, 23, 1]. We extracted from these works the main characteristics they exploit and grouped these characteristics according to the type of information modeled, defining a new taxonomy. Then, all selected works were classified according to the defined groups. To evaluate these groups, we derived a descending ordered list of Top-50 influencers, according to the scores defined by each strategy. Comparing these lists, we confirmed that lists derived from strategies belonging to the same class present the highest level of agreement. Moreover, using Principal Component Analysis (PCA), we analyzed the complementarity of information modeled by strategies belonging to the same class, demonstrating that strategies belonging to the same class are strongly correlated. Further, we used PCA as a meta-learning process to combine linearly distinct strategies, alleviating the difficulty on deciding which strategy researchers would adopt. As future work, we intend to extent the proposed taxonomy by inspecting other strategies existing in the literature. The goal is to combine all of them using PCA, achieving a new metric more complete w.r.t the characteristics exploited.

References

[1] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proc. 4th ACM WSDM*, 2011.

[2] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201, 2001.

[3] Phil E Brown and Junlan Feng. Measuring user influence on twitter using modified k-shell decomposition. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 2010.

[5] Chandler Davis. The norm of the schur product operation. *Numerische Mathematik*, 1962.

[6] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. In *Proc. 14th ACM-SIAM*, 2003.

[7] Andrea Galeotti and Sanjeev Goyal. The law of the few. *The American Economic Review*, 2010.

[8] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[9] M. U. Ilyas and H. Radha. Identifying influential nodes in online social networks using principal component centrality. In *Communications (ICC), IEEE International Conference*. IEEE, 2011.

[10] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 2010.

[11] Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010.

[12] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proc. of the 13th ACM SIGKDD*, 2007.

[13] Yung-Ming Li, Chia-Hao Lin, and Cheng-Yang Lai. Identifying influential reviewers for word-of-mouth marketing. *Electron. Commer. Rec. Appl.*, 9(4):294–304, jul 2010.

[14] Frank McCown and Michael L. Nelson. Agreeing to disagree: Search engines and their public interfaces. In *Proc. 7th ACM/IEEE-CS JCDL*. ACM, 2007.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[16] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901.

[17] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[18] Arlei Silva, Sara Guimarães, Wagner Meira, Jr., and Mohammed Zaki. Profilerank: Finding relevant content and influential users based on information diffusion. In *Proc. of the 7th SNA-KDD*. ACM, 2013.

[19] Karthik Subbian, Dhruv Sharma, Zhen Wen, and Jaideep Srivastava. Finding influencers in networks using social capital. In *Proc. of IEEE/ACM ASONAM*, 2013.

[20] Thomas W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 1996.

[21] Thomas W. Valente and Patchareeya Pumpuang. Identifying Opinion Leaders to Promote Behavior Change. *Health Education & Behavior*, 2007.

[22] Christophe Van den Bulte and Yogesh V. Joshi. New product diffusion with influentials and imitators. *Marketing Science*, 2007.

[23] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *Proc. the 20th WWW*, 2011.