

## A Computational Approach to Measuring the Correlation between Expertise and Social Media Influence for Celebrities on Microblogs

Wayne Xin Zhao, Jing Liu, Yulan He,  
Chin-Yew Lin and Ji-Rong Wen

Received: date / Accepted: date

**Abstract** *Social media influence* analysis, sometimes also called authority detection, aims to rank users based on their influence scores in social media. Existing approaches of social influence analysis usually focus on how to develop effective algorithms to quantize users' influence scores. They rarely consider a person's expertise levels which are arguably important to influence measures.

In this paper, we propose a computational approach to measuring the correlation between expertise and social media influence, and we take a new perspective to understand social media influence by incorporating expertise into influence analysis. We carefully constructed a large dataset of 13,684 Chinese celebrities from Sina Weibo (literally "Sina microblogging"). We found that there is a strong correlation between expertise levels and social media influence scores. Our analysis gave a good explanation of the phenomenon of "top across-domain influencers". In addition, different expertise levels showed influence variation patterns: e.g., (1) high-expertise celebrities have stronger influence on the "audience" in their expertise domains; (2) expertise seems to be more important than relevance and participation for social media influence; (3) the audiences of top expertise celebrities are more likely to forward tweets on topics outside the expertise domains from high-expertise celebrities.

**Keywords** Social media influence · Expertise · Microblog

---

Xin Zhao (contact author), Ji-Rong Wen  
School of Information, Renmin University of China, Beijing, P.R.China  
E-mail: batmanfly@gmail.com, jirong.wen@gmail.com

Jing Liu, Chin-Yew Lin  
Microsoft Research, Beijing, P.R.China  
E-mail: legendarydan@gmail.com, cyl@microsoft.com

Yulan He  
School of Engineering and Applied Science, Aston University, UK  
E-mail: y.he@cantab.net

## 1 Introduction

In recent years, with the rapid growth of social media platforms such as Twitter and Facebook, social network analysis has become a popular research topic. Specifically, considerable research has focused on understanding and analyzing social media influence [11, 25, 20]. The best known of the social media influence measurement tools, Klout<sup>1</sup>, measures one's overall online influence by combining Twitter data such as follower count, retweets, unique mentions, how influential the people that retweet are etc., with Facebook data such as comments, likes, and the number of friends in one's network. Nevertheless, when measuring a user's influence on social media, existing approaches rarely take into account of her expertise area and hence make no distinction between her influence on the topics relating to her expertise area and that on topics outside of her expertise area. Furthermore for users in the same expertise area, no distinction was made between influence from users with different expertise levels.

Why is it important to consider expertise when performing social media influence analysis? Social influence has already been shown to be closely related to expertise in various application areas. In Community-based Question-Answering (CQA) portals, askers would be more likely to adopt the answer from a user with high-expertise [38]; in scientific publications, papers from domain experts [13] are more likely to be cited by other researchers; in online product recommendation, "human experts" are perceived to be more persuasive for product adoption than other consumers [23]. These examples suggest that expertise is an important factor in social influence analysis.

Instead of analyzing all types of users in the social web, we are particularly interested in celebrities. As an attractive merit, some online social networking websites adopt a "verified user system" to highlight the importance of opinion leaders. Indeed, most of these verified users are celebrities from business, education, media and entertainment industries. Celebrities can use their own reputation, by virtue of the privileges provided by the site, quickly win a lot of attentions. As such, business organizations often use celebrity endorsements to promote their products or services. Important factors to consider when using celebrities to endorse products are their relevance and influence. For example, given a product, should we select a celebrity in an irrelevant domain but with significant influence globally or the one in a relevant domain but with a relatively weaker influence globally? Without considering the expertise of celebrities in influence analysis, we cannot get a deep understanding of how celebrities exert their influence on the public and how their influence varies across different topics or domains.

The above suggests that considering expertise for social media analysis would help us gain deep insights into social influence, leading to more effective findings to recommender systems and online advertising systems. Indeed, expertise and social media influence provide two different perspectives to view a celebrity: either by the expertise level recognized within her expertise area or by the degree of her influence exerted on the public (i.e. common users) in social media. Understanding the correlation between two perspectives has served as a fundamental question implicitly in sociology and social philosophy [21]. Analyzing relationship between expertise and

---

<sup>1</sup> <http://klout.com>

social influence could shed some light on understanding the effects of expertise on social influence. Nevertheless, prior empirical studies on examining the relations between expertise and social influence are typically conducted on small datasets. In this paper, we present the first study which takes into account of expertise into social influence analysis on a large dataset constructed from Sina Weibo<sup>2</sup>, which is the largest Chinese microblog service, and we propose a computational approach to measuring the correlation between expertise and social media influence.

To perform such an analysis, we carefully construct a dataset of 13,862 celebrities on Sina Weibo from four domains. We distinguish the expertise category from non-expertise category of celebrities, and carry out the correlation analysis *within an expertise domain* and *across domains*. Our work provides several important findings to understand social influence of celebrities in online social media:

- There is a strong correlation between expertise level and social media influence on both expertise categories and non-expertise categories. Although the overall correlation patterns are similar, the correlation degrees measured in the expertise-related categories are much more significant than those measured on non-expertise categories. This is due to the phenomenon that users are more likely to forward the tweets of topics relating to the celebrities' expertise domains than those outside their expertise domains.
- Top-expertise celebrities are likely to influence across domains and even more influential than the in-domain celebrities. Comparing to the audiences of lower expertise celebrities, the audiences of top expertise celebrities are more likely to forward tweets on topics outside the expertise domains from high-expertise celebrities.
- Our work has close connection with related studies in sociology, and further shows important new findings: expertise seems to be more important than relevance and participation for social media influence.

The rest of the paper is organized as follows. The overview of our work is presented in Section 2, followed by our data preparation in Section 3. Section 4 studies how to measure expertise levels and influence scores. Section 5 discusses the detailed results of social influence analysis by incorporating expertise, and shows insights drawn from the analysis of the data collection. Section 6 describes our findings in connection to the Status Characteristic Theory. Section 7 presents related work. Finally, Section 8 concludes the paper.

## 2 Overview of the Methodology

The overview of the methodology is listed by sections as follows:

- Section 3 discussed the data preparation. We crawl the detailed information of 13,862 celebrities on Sina Weibo and obtain their domain labels from an existing influence system, Bing Influence<sup>3</sup>. Furthermore, as celebrities can publish

---

<sup>2</sup> <http://weibo.com/>

<sup>3</sup> <http://cn.bing.com/yixiangli/>

tweets on any topics not necessarily relating to their own domain, for example, a movie star from the *Entertainment* domain can comment on a news event, we analyze contents published by the celebrities by mapping the topics discussed into nine content categories. As will be shown later, such content categories include additional five categories such as *Education, News, Life*, etc., apart from the four pre-labelled celebrity domains mentioned above (see Table 5). Content categorization allows us to compare a celebrity’s influence on topics relating to her expertise domain and topics outside of her expertise domain. We classify each tweet of celebrities into these nine categories by using topic modeling methods.

- Section 4 presented the expertise and influence measurement. By following the previous study [9], we propose to use the number of times the tweets published by a celebrity have been retweeted to measure her social media influence, either by the total volume or by average. We hypothesize that the “following” links between celebrities in the same domain is an acknowledgement of expertise, and propose to use a domain-specific PageRank score to measure expertise levels by leveraging the following links between celebrities. This enables us to study the differences of influence patterns of celebrities with different expertise levels in the same expertise domain.
  - Section 5 studied the overall within-domain correlation between expertise and social media influence. With the defined expertise and influence measurement, we can derive expertise ranking and influence ranking (either expertise based or non-expertise based). Furthermore, we proposed to use Spearman’s rank correlation coefficient as the correlation measurement. There is a strong correlation between expertise level and social media influence on both expertise categories and non-expertise categories, but the correlation degrees measured in the expertise-related categories are much more significant than those measured on non-expertise categories. We explained this finding by investigating into the forwarding behaviours of users.
  - Section 6 further analyzed the across-domain influence of top-expertise celebrities, since that high-expertise celebrities are more influential in their own domains. We introduced a new perspective to explain why high-expertise are more likely to influence across domains. We try to reveal the differences between high-expertise and low-expertise celebrities by analyzing the propagators of celebrities.
- For convenience, we summarize the notations used throughout the paper in Table 1.

### 3 Data Preparation

In this section, we discuss how to construct the data collection for our influence analysis.

#### 3.1 Celebrity Dataset Collection

We obtained a list of celebrities from Bing Influence. Bing Influence measures the relative influence of Chinese celebrities in four categories or domains *Entertainment*,

**Table 1** Notations.

Notations	Explanations
$c$	A domain or category
$w$	A word
$d$	A tweet
$t$	A topic
$u, v$	A user
$c_u$	The domain of user $u$
$e_u$	The expertise level of $u$
$\rho_c^u$	The influence of $u$ on a content category $c$
$\mathcal{C}$	The set of domains or categories
$\mathcal{D}$	The set of tweets
$\mathcal{D}^c$	The set of tweets published by the celebrities from domain $c$
$\mathcal{D}_{(E)}^c$	The set of tweets on the expertise content categories, $\mathcal{D}_{(E)}^c \subset \mathcal{D}^c$
$\mathcal{D}_{(NE)}^c$	The set of tweets on the non-expertise content categories, $\mathcal{D}_{(NE)}^c \subset \mathcal{D}^c$
$\mathcal{T}$	The set of topics
$\mathcal{T}_c$	The set of topics in category $c$
$\mathcal{F}$	The set of propagators
$\mathcal{F}_{\geq n}$	The set of propagators with at least $n$ retweets
$\mathcal{F}_{(E)}^c$	The set of propagators who have made retweets on tweets in $\mathcal{D}_{(E)}^c$
$\mathcal{F}_{(NE)}^c$	The set of propagators who have made retweets $\mathcal{D}_{(NE)}^c$
$\mathcal{R}$	The set of retweets
$\mathcal{R}_{(E)}^c$	The set of retweets made on expertise tweets in $\mathcal{D}_{(E)}^c$
$\mathcal{R}_{(NE)}^c$	The set of retweets made on non-expertise tweets in $\mathcal{D}_{(NE)}^c$

**Table 2** The total number of users per domain and average followings per user by domains.

Domains	#users	→Enter.	→Scien.	→Busi.	→Sports
<b>Entertainment</b>	8,586	<b>64.1</b>	3.4	5.1	1.9
<b>ScienTech</b>	1,801	22.8*	<b>52.7</b>	26.3	1.5
<b>Business</b>	2,024	23.6	16.7	<b>42.3</b>	1.6
<b>Sports</b>	1,451	16.1	1.0	2.0	<b>25.1</b>

Explanation: This entry of “22.8” denotes that a celebrity in ScienTech has followed 22.8 celebrities in Entertainment on average.

**Table 3** Detailed statistics of #followers by domains.

Domains	#users	minimum	maximum	median	average
<b>Entertainment</b>	8,586	10,069	45,199,689	43,760	376,692
<b>ScienTech</b>	1,801	10,082	38,522,520	29,125	157,185
<b>Business</b>	2,024	10,104	15,166,905	38,727	251,693
<b>Sports</b>	1,451	10,724	21,351,088	48,515	328,226

*ScienTech*, *Business* and *Sports*. Each celebrity is manually classified into a single domain according to their professions and has a global (shared across all the domains) and comparable influence score calculated with huge amounts of data from multiple social media signals. We only use the domain labels from Bing Influence but not the influence scores because the scores are not domain-specific. We select Sina Weibo as the microblog platform, which is the largest Chinese microblog service. We start with these celebrities’ names and crawl their tweets and following/follower informa-

**Table 4** Data statistics after preprocessing.

#users	#tweets	#tokens	#terms	Starttime	Endtime
13,862	26,725,958	406,128,627	80,550	April, 2011	April, 2013

tion from Sina Weibo if their account names can be found.<sup>4</sup> For each tweet, we can obtain the total number of being retweeted, and we use this statistics to measure social media influence as will be described later. The time span of the crawled tweets is from April 2011 to April 2013. We have collected 13,864 celebrities with a total of 26,725,958 tweets. We perform a few standard text preprocessing steps on Weibo text data including segmentation, stopword removal and low-frequency ( $< 50$ ) term removal. The data statistics are summarized in Table 4.

Table 2 shows the statistics of each domain and the average followings among celebrities. We can see that the *Entertainment* domain has the largest proportion of the entire population, consisting of 61% of total celebrities. We also notice that celebrities tend to follow other celebrities in their own domain, which shows a strong evidence of “homophily” [15, 33]. Table 3 shows the statics of celebrities by domains. We can see celebrities in all the four domains have a similar minimum number of followers which is over 10,000. *Entertainment* domain has the largest proportion of the entire population, consisting of 61% of total celebrities, and it has the largest maximum and average number of followers compared to the other three domains. When counting the celebrities who have more than 1 million followers, *Entertainment* has the most number of such celebrities among the four domains.

### 3.2 Tweet Content Categorization

Since our aim is to leverage tweet content in influence analysis, we describe how to extract topics from tweets published by celebrities and how to group the extracted topics into different content categories.

A *topic* is a subject discussed in one or more documents. Examples of topics include news events such as “US government shutdown” and long-standing subjects such as “music” and “movie”. Because our dataset is large, it is only feasible to use fully unsupervised or weakly supervised methods to automatically discover topics. Here we adopt LDA-based topic modeling to perform topic discovery. Each topic is assumed to be represented by a multinomial distribution of words. As suggested by a previous study [37], we aggregate all the tweets of a celebrity as a document and run a modified LDA which incorporates background models to reduce the affect of background words with a topic number set to 100.<sup>5</sup>

A *content category* groups topics belonging to a common subject area together. From manual inspection of the top ten topic words in the list of 100 topics, we identified nine content categories where the first four categories are the same as the domains of celebrities from Bing Influence. We manually assigned a topic to one of these nine categories and discarded topics that lack of coherent semantics. Finally 87 topics are

<sup>4</sup> Sina Weibo has verified most of these celebrities, thus we can simply match their actual names with account names.

<sup>5</sup> We have tried other numbers, but found 100 was the optimal choice for our dataset in terms of topic interpretation.

assigned to one of the nine categories. The topic categorization results were discussed with the experts from the Bing Influence group to ensure their validity. The nine content categories together with their definitions are given in Table 5. We notice that *Entertainment* is the top category with the most number of topics. It is perhaps not surprising since the *Entertainment* domain consists of the largest proportion, i.e. 61% of celebrities in our dataset.

**Table 5** Definitions of content categories and the corresponding number of topics for each category. Categories correspond to celebrity domains are marked in bold.

Category	#topics	Definition
<b>Entertainment</b>	29	Television, Movies, Music, Media, Radio, Theater, Fashion&Style
<b>ScienTech</b>	4	IT technology, applications, platforms
<b>Business</b>	6	sales, marketing, economic, shop, consulting, stock
<b>Sports</b>	4	athletes, sports teams, leagues, sports events, or kinds of sports, including basketball, baseball, soccer, etc.
Education	10	books, history, lessons, universities, arts, forums
News	7	events or topics in China and in the world
Life	15	food, drink, health, fitness, medicine, travel, airline, environment, charity, Internet games, lifestyle, religion, pets and quotes
Chat&Status	9	chat idioms and gossip
Weibo	3	Sina Weibo related topics, e.g., functions settings and conventions

With the assignments of topics to categories, we can automatically classify each tweet to one of these nine categories. Formally, denoting the conditional probability of word  $w$  given a topic  $t$  by  $P(w|t)$ , a tweet  $d$  is categorized to category  $c_d$  as follows

$$c_d = \arg \max_{c \in \mathcal{C}} \left\{ \arg \max_{t \in \mathcal{T}_c} P(d|t) \right\}, \quad (1)$$

$$= \arg \max_{c \in \mathcal{C}} \left\{ \arg \max_{t \in \mathcal{T}_c} \prod_{w \in d} P(w|t) \right\}. \quad (2)$$

Note that with the above definition, we have nine content categories and only four of them correspond to celebrity domains. Given a celebrity, the content category which is the same as her expertise domain will be referred to as her **expertise category** while the other content categories will be referred to as **non-expertise categories**.

To help understand our work, we have implemented a demonstration system, named as *WeiboRank*, which presents the built categories, the extracted topics within each category and top-influential celebrities ranked with our social media influence measurement (see more in Section 4) by topics<sup>6</sup>. The start page is a category index page which group topics by categories, and we can click the category tabs on the left to switch between categories. Category-specific topics are shown in word clouds. We do not consider the “Weibo” category in Table 5, since it is generally related to the

<sup>6</sup> The system link is <http://playbigdata.com/batmanfly/weiborank>.

**Table 6** Number of users who have re-tweeted at least 1, 10 or 100 tweets from the celebrities in our dataset.

$\mathcal{F}_{>1}$	$\mathcal{F}_{>10}$	$\mathcal{F}_{>100}$
3,477,506	131,626	5,055

Weibo service and does not contain topical semantics. We can further click on the topical word clouds, and it will lead to the topic index page. In the topic index page, the topical word cloud is shown on the left, and the top-influential celebrity ranking is shown on the right. By clicking on a celebrity, we can get her topic-specific keywords.

### 3.3 Propagator Dataset Collection

To better analyze how celebrities exert their influences on their followers, we further crawl the data from users who have re-tweeted tweets published by celebrities. Due to the fact that these celebrities attract hundreds of billions of retweets, it is infeasible to crawl all the retweets within the access limit of Sina Weibo. As such, we ignore the retweets by the users with fewer than ten friends links and ten tweets, and obtain a total of 10,262,482 retweets from 3,477,506 unique users. We consider three types of users based on their forwarding frequencies: users with at least one retweet, users with at least ten retweets and users with at least a hundred retweets, denoted by  $\mathcal{F}_{\geq 1}$ ,  $\mathcal{F}_{\geq 10}$  and  $\mathcal{F}_{\geq 100}$  respectively. Different from previous studies, here we only consider propagators as direct “audience” of celebrities instead of all the followers because their re-tweeting activities explicitly show that they are somehow influenced by the celebrities.

## 4 Expertise and Influence Measurement

In this section, we first describe how to measure a celebrity’s expertise level in one particular *domain*. Then, we present our method on measuring social media influence in a *content category*.

### 4.1 Measuring Individual Expertise Level

*Expertise* consists of those characteristics, skills and knowledge of a person (that is, expert) or of a system, which distinguish experts from novices and less experienced people<sup>7</sup>. Although a person can be experts in multiple domains, in this paper, we want to identify the primary expertise of a celebrity, i.e., a single domain corresponding to her profession or occupation.

<sup>7</sup> <http://en.wikipedia.org/wiki/Expert>

Expertise measurement has been a very important research topic and receives much attention from research communities. Generally speaking, on online social networks, there can be two ways to measure expertise of celebrities: one way is to consider the acknowledgement or evaluation from domain-specific experts or members (similar to expertise evaluation in enterprises), where expertise is likely to correspond to the domain status [27]; while the other is to estimate expertise from the response of common users (e.g. the number of best answers received from online answer-questioning forums). Although the former way is more consistent with the expert evaluation in reality, it is usually difficult to capture direct acknowledgement signals from domain experts. Fortunately, the increasing popularity of online social network platforms enables use of, for example, the following links between two users, as acknowledgement of expertise for expertise learning [29]. However, another challenge is how to identify the expertise domain of a user. The previous study [29] is to simply learn the acknowledgement of expertise from all users not limited to a specific expertise-related domain. Our solution is to 1) leverage the available domain labels of celebrities to identify their expertise areas; and 2) make use of the *following* link for capturing the acknowledgement of expertise.

Formally, we assume that there is a set of domains  $\mathcal{C}$  and a set of celebrities  $\mathcal{U}$  on the data collection we consider. Let  $u$  denote a celebrity user, and  $c$  denote a domain. Each celebrity  $u \in \mathcal{U}$  is an expert at **only one domain**  $c_u \in \mathcal{C}$ , and  $c_u$  is called as the expertise domain of  $u$ . We further introduce a real variable  $e_u$  to measure the expertise level of  $u$ , where  $0 \leq e_u < 1$ . We propose to use a domain-specific PageRank score to measure the expertise level of a celebrity. Recall that each celebrity is labeled with one of the four domain labels from Bing Influence and we have the “following” links between celebrities. Given two celebrities  $u$  and  $v$  in the same domain, we consider a following link from  $u$  to  $v$  to be  $u$ ’s acknowledgement of  $v$ ’s expertise. Specially, we build the following graph among celebrities in the same domain. Since our assumption is based on the acknowledgement from within-domain experts, for each domain, we only consider **celebrities from this domain and links between them**, and run the standard PageRank algorithm to obtain the PageRank scores as expertise level scores, i.e., given a category  $c$ ,  $\forall u \in \mathcal{U}_c, e_u = \text{PageRank}_c(u)$ . To better study the impact of expertise level in later experiments, we sort celebrities within a domain  $c$  by their expertise levels in a descending order and split them into ten equal bins with the same number of celebrities, denoted by  $\mathcal{U}_{c,i}$  where  $i = 1, \dots, 10$  and a smaller value of  $i$  indicates this group has a higher expertise level. We do not split celebrities by setting PageRank score ranges, which are likely to result in bins of varying sizes. Keeping bins with the same size would make the analysis easier to follow and reduce the effect of sample sizes on the statistics within each bin. Our definition of expertise level is closely related to the expert status in domains: those team members who are identified as possessing expertise are often afforded power and status [10].

## 4.2 Measuring Social Media Influence

Our definition of social media influence aims to measure the influence that celebrities exert on the public in a specific *content category* on microblogs. Note that social media influence can be understood as the instantiation of social influence on online social media, which is not necessarily equal to social influence in reality. We denote the influence of  $u$  on a content category  $c$  by a real value  $\rho_c^u$ , where  $\rho_c^u \geq 0$ . How to measure social media influence has been an important research issue in recent years. Instead of adopting complex metrics, we use the *retweet* mechanism as the signal to measure the influence based on the assumptions: 1) A celebrity has influenced a common user if the user has forwarded a tweet of the celebrity; and 2) The more tweets that common users have forwarded from a celebrity, the more influential the celebrity is.

With the above assumptions, we propose two social media influence measures. The first measure is the categorical ***total influence***. The total influence of a celebrity  $u$  on a category  $c$  is defined as the total number of times that her tweets with a content category label  $c$  have been retweeted. Total influence measures the overall influence degree that a celebrity has over the public on microblogs.

The second measure we introduce is the categorical ***average influence***. The average influence of a celebrity  $u$  on a category  $c$  is defined as the average number of times that her tweets with a content category label  $c$  have been retweeted.

Both influence measures can contain categorical influence bias, i.e. tweets in some categories will receive more retweets than others. E.g. tweets related to “breaking news” tend to receive significantly more retweets than others. In order to reduce the categorical bias, for both influence measurements, we divide the statistics of the retweet numbers by the maximum retweets per tweet receives in a category.<sup>8</sup>

We acknowledge that the metrics we used to quantify user influence are simple. However, retweeting someone else’s content signals an endorsement on one’s posts and thus the number of retweets naturally represents a measure of social success. There might be many other factors contribute to the measurement of social influence such as the activity level (number of tweets published), the audience size (the number of followers), the size of the retweet cascades etc. Nevertheless, we aim to measure influence in different content categories in order to distinguish between influence in the expertise-related categories and that in non-expertise categories. As such, some factors such as the audience size are not applicable here. Also, due to the access limit imposed by Sina Weibo, it is not possible for us to trace down the complete retweet cascades. Hence, more complicated influence measures such as those based on the size of propagation cascades [3] cannot be easily implemented.

### *Summary*

In this paper, a celebrity only has one associated expertise domain corresponding to her profession and her expertise level is measured by a real value which is comparable

<sup>8</sup> Although we cannot obtain the entire set of retweets of a tweet, Sina API provides the exact number of being retweeted for each tweet. Thus, the statistics of retweet numbers are accurate and not based on incomplete propagator sets in Table 6.

among celebrities in the same domain. A celebrity can publish tweets with content either relating to her expertise domain (i.e. expertise category) or outside her expertise domain (i.e. non-expertise categories). Thus, a celebrity can be potentially influential over more than one content category and social media influence is thus measured by a vector of positive values, each of which indicates the influence strength on the corresponding content category.

In what follows, we perform the analysis in two aspects: (1) within-domain and (2) across-domain correlations between expertise and social media influence.

## 5 Measuring Within-Domain Correlations between Expertise and Social Media Influence

In sociology, there is a common belief that expertise is related to social influence or power [10]. But there is a lack of validation of such a correlation on large datasets [21] and no prior studies presented a quantitative measurement of the correlation. Without a formal conclusion on the correlation, it is difficult to explain why the involvement of expertise works and how it affects the performance in some tasks. To address the above problem, we first formally test the within-domain correlation between expertise and social media influence.

### 5.1 Correlation significance test

We adopt the widely used Spearman’s rank correlation coefficient as the measurement. The Spearman correlation coefficient<sup>9</sup> is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size  $n$ , the  $n$  raw scores  $X_i, Y_i$  are converted to ranks  $x_i, y_i$ , and the Spearman correlation coefficient  $\rho$  is computed as  $r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2-1)}$ , where  $d_i = x_i - y_i$ . It can deal with *duplicate values*. The Spearman’s coefficient lies in the interval  $[-1, 1]$ , and a value of “+1” or “-1” indicates that a perfect, positive or negative Spearman correlation. When  $n$  is small, it is easy to test the correlation significance by looking up the Spearman significance table; when  $n$  is large,  $r_s \sqrt{n-1}$  is approximately  $\mathcal{N}(0, 1)$ , thus we can perform single-tail significance test based on the normal distribution [36, 1].

The main reason to adopt rank correlation measures is that it considers the relative order between items instead of the absolute values attached to each item. Thus, the effect of metrics from different rankings on the correlation values is relatively small. The Spearman rank correlation is used instead of the Pearson correlation method because Pearson correlation measures the strength of the linear relationship between normally distributed data while Spearman rank correlation makes no assumptions about the distribution of the data. It is therefore more appropriate to use Pearson correlation for our data here which may have large outliers that hide meaningful relationships between them.

The input rankings for correlation computation are 1) the expertise ranking and 2) a categorical influence ranking of a content category. We have two kinds of influence

<sup>9</sup> [http://en.wikipedia.org/wiki/Spearman's\\_rank\\_correlation\\_coefficient](http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient)

measurements, namely total influence and average influence. For each influence measurement, we can obtain an influence ranking on one of the nine content categories. We refer to the influence ranking on the expertise category as *expertise influence ranking*, while influence rankings on other categories as *non-expertise influence rankings*. For each celebrity domain, we first obtain the expertise ranking of celebrities using the domain-specific PageRank, then we can obtain the corresponding influence rankings of these celebrities in nine content categories: one is the expertise influence ranking and others are non-expertise influence rankings. Given an expertise ranking and an influence ranking, we can compute their Spearman’s rank correlation coefficient. **It is worth noting that celebrities from a specific domain have only one expertise ranking, but have more than one influence ranking with one in each content category (recall that we have a total of 9 content categories as listed in Table 2).**

## 5.2 The overall observations

We use the Spearman’s rank correlation coefficient to measure the overall correlation degree between the expertise ranking and an influence ranking. We present the results in Table 7. From this table, we have the following observations:

1) For both influence measures, expertise ranking is positively correlated with influence ranking. Nearly all the entries of expertise categories have a value larger than 0.5. Note that our correlation analysis is based on large samples, i.e.  $n \geq 1000$ , hence a value larger than 0.5 indicates very strong correlation between expertise and social media influence.

2) The correlation value of expertise influence ranking is significantly and consistently larger than that in non-expertise influence ranking for four domains. This can be easily verified by using the average correlation values together with the standard deviations provided in Table 7.

3) For total influence, the correlation values of expertise entries become more significant but the values of non-expertise entries become less significant.

We first analyze the first observation. Given a domain, a large correlation value between an expertise ranking and an influence ranking indicates that the influence of a celebrity in a content category largely depends on her own domain-specific expertise level. It is intuitive that the influence of a celebrity in the expertise content category depends on her own expertise level. Interestingly, the evidence still holds across different non-expertise content categories: There are still strong correlations between expertise and influence measured in non-expertise content categories although the degree of correlation varies in different expertise domains. In social philosophy, it has been widely acknowledged that status and relevance are two most important factors which determine the individual influence or power [21]. In our current study, expertise levels can be understood as a measure of domain status, and our finding indicates that status (expertise level) is strongly correlated with celebrities’ influence across different content categories, even in less relevant (non-expertise) content categories, *which indicates that expertise is more important than relevance for social media influence.*

**Table 7** Overall influence correlation comparison for expertise and non-expertise categories by four expertise domains. For eight non-expertise categories, we present their mean and the standard deviation. All the correlation values are significant at the level of 0.01. The expertise entries are consistently larger than the non-expertise entries.

Domains	Average Influence		Total Influence	
	Expertise	Non-Expertise	Expertise	Non-Expertise
Entertainment	0.61	0.56±0.04	0.65	0.45±0.09
ScienTech	0.56	0.48±0.04	0.62	0.40±0.09
Business	0.61	0.56±0.02	0.66	0.51±0.06
Sports	0.50	0.48±0.04	0.51	0.42±0.08

The last two observations have revealed the difference between the expertise and non-expertise categories: the expertise ranking has a stronger correlation with expertise content categories than non-expertise categories. We next analyze why such a difference exists by investigating into celebrities’ audiences.

### 5.3 Empirical analysis of the observations based on “audience” of celebrities

In the above, we have found that expertise level has a stronger correlation with social media influence on expertise category than non-expertise categories. It is worth finding out what the differences in influence are between expertise category and non-expertise categories for celebrities in the same domain. The entire analysis is constructed in two aspects.

First, we compute the statistics about the average number of retweets received per tweet on four expertise domains. Table 8 shows the average number of retweets received per tweet on four expertise domains. On average, we can see that a celebrity receives more retweets on a tweet from her own expertise category, which indicates that celebrities are indeed more influential on the corresponding expertise categories.

**Table 8** Average number of retweets received per tweet on four expertise domains. We consider two types of content categories: expertise categories and non-expertise categories. We mark the standard error of the average in parentheses.

Domain	Expertise	Non-expertise
Expertise	<b>380.94</b> (6.23)	374.44 (6.38)
Scien.	<b>226.12</b> (4.70)	220.28 (3.70)
Busi.	<b>225.00</b> (4.40)	224.61 (3.25)
Sports	<b>127.53</b> (3.61)	120.13 (2.03)

Second, since we measure social media influence through the forwarding behaviors of ordinary users on microblogs, our main idea is to explain our observations by gaining insights into the propagators of these celebrities, i.e. their “audience”. Generally speaking, a celebrity can write tweets either on her own expertise category or non-expertise categories, and these tweets might be forwarded by her followers. Here we want to analyze how differently these propagators behave when they retweet in these two different types of content categories. Formally, given an expertise domain

$c$ , let  $\mathcal{D}^c$  denote the set of tweets published by the celebrities from domain  $c$ . Furthermore, we incorporate subscripts into  $\mathcal{D}^c$  to distinguish two types of tweets: 1)  $\mathcal{D}_{(E)}^c$  denotes the set of tweets on domain  $c$ 's corresponding expertise content category; 2)  $\mathcal{D}_{(NE)}^c$  denotes the set of tweets on the non-expertise content categories. For  $\mathcal{D}_{(E)}^c$ , let  $\mathcal{F}_{(E)}^c$  denote the set of users who have forwarded at least a tweet in  $\mathcal{D}_{(E)}^c$  and  $\mathcal{R}_{(E)}^c$  denote the set of tweets that have been retweeted in  $\mathcal{D}_{(E)}^c$  by these propagators in  $\mathcal{F}_{(E)}^c$ . Similarly, we can define  $\mathcal{F}_{(NE)}^c$  and  $\mathcal{R}_{(NE)}^c$ . Then given an expertise domain  $c$  and its set of celebrities, we introduce two measures to calculate: (1) the average number of tweets that have been retweeted by a user in the *expertise tweet set* (i.e.  $\mathcal{D}_{(E)}^c$ ); (2) the average number of tweets that have been retweeted by a user in the *non-expertise tweet set* (i.e.  $\mathcal{D}_{(NE)}^c$ ) respectively:  $n_{(E)}^c = \frac{|\mathcal{R}_{(E)}^c|}{|\mathcal{F}_{(E)}^c|}$  and  $n_{(NE)}^c = \frac{|\mathcal{R}_{(NE)}^c|}{|\mathcal{F}_{(NE)}^c|}$ .

We consider three propagator sets which have different degrees of forwarding activities as have been previously shown in Table 6. We present the results in Table 9 and mark values in the expertise category in bold. Interestingly, we can see that celebrities' tweets on the expertise category receive more retweets from per propagator on average. In other words, a propagator is more likely to forward tweets from a celebrity's expertise category than her non-expertise categories. This indicates that celebrities are more "attractive" or "influential" in their expertise categories and thus receive more attentions from their propagators. Interestingly, the difference between  $n_{(E)}^c$  and  $n_{(NE)}^c$  enlarges when we consider more active propagators. We may conclude that celebrities have stronger influence in their expertise related domains.

**Table 9** Average number of retweets per propagator on four expertise domains. We consider two types of content categories: expertise categories and non-expertise categories. We mark the standard error of the average in parentheses.

Celebrity Domain	$\mathcal{F}_{>1}$		$\mathcal{F}_{>10}$		$\mathcal{F}_{>100}$	
	$n_{(E)}^c$	$n_{(NE)}^c$	$n_{(E)}^c$	$n_{(NE)}^c$	$n_{(E)}^c$	$n_{(NE)}^c$
Expertise	<b>1.37</b> (0.001)	1.11 (0.005)	<b>2.12</b> (0.013)	1.51 (0.024)	<b>6.52</b> (0.587)	2.66 (0.353)
Scien.	<b>1.47</b> (0.001)	1.19 (0.004)	<b>1.88</b> (0.006)	1.49 (0.016)	<b>4.90</b> (0.191)	3.49 (0.318)
Busi.	<b>1.42</b> (0.001)	1.22 (0.005)	<b>1.75</b> (0.006)	1.32 (0.014)	<b>5.49</b> (0.131)	2.32 (0.241)
Sports	<b>1.14</b> (0.001)	1.09 (0.005)	<b>1.31</b> (0.006)	1.19 (0.017)	<b>1.55</b> (0.109)	1.29 (0.161)

#### 5.4 Correlation value comparison to other measures

Given our influence measurements, it is intuitive to consider two other potentially correlating factors: 1) activity level of a celebrity; 2) audience size of a celebrity. In this part, we would like to examine how these two factors are correlated with influence measurements and in comparison with expertise, which one has stronger correlation with social media influence. We measure the activity level by the number of tweets a celebrity has published (*#tweets*), and measure the audience size by the number of followers of a celebrity (*#followers*). Given a domain, we can generate a descend-

ing ranking of celebrities using either *#tweets* or *#followers*, and then compute the correlation values between the generated ranking and influence rankings.

We present the correlation values of *#tweets* and *#followers* respectively in Table 10 and 11. Overall, we can see that 1) *#tweets* has a good correlation with total influence but a very poor correlation with average influence. This is perhaps not surprising since publishing more tweets are likely to have more tweets being retweeted. 2) *#followers* has a good correlation average influence but relatively a weak correlation with total influence.

**Table 10** Overall influence correlation comparison for expertise and non-expertise categories by *#tweets*.

Domains	Average Influence		Total Influence	
	Expertise	Non-Expertise	Expertise	Non-Expertise
Entertainment	0.14	0.05±0.14	0.40	0.51±0.08
ScienTech	0.04	0.11±0.07	0.53	0.65±0.05
Business	0.02	0.15±0.08	0.57	0.65±0.03
Sports	0.05	0.18±0.15	0.58	0.66±0.03

**Table 11** Overall influence correlation comparison for expertise and non-expertise categories by *#followers*.

Domains	Average Influence		Total Influence	
	Expertise	Non-Expertise	Expertise	Non-Expertise
Entertainment	0.58	0.53±0.05	0.53	0.44±0.09
ScienTech	0.45	0.46±0.03	0.38	0.39±0.03
Business	0.51	0.49±0.04	0.42	0.38±0.05
Sports	0.48	0.44±0.06	0.37	0.34±0.06

Now we take our expertise measurement into comparison by cross comparing Table 7, 10 and 11. We can clearly observe that *expertise* has stronger correlation values in expertise-related categories with both influence measurements compared to *#tweets* and *#followers*. Although *expertise* has lower correlation values in non-expertise entries with total influence compared to *#tweets*, our *expertise* has the discriminativity power between expertise and non-expertise categories. It can yield larger correlation values in expertise categories than non-expertise categories, while both *#tweets* and *#followers* cannot capture such patterns. Overall, our proposed *expertise* measurement is better to capture the correlation with both influence measurements and yield more consistent results with our intuition. To better understand the relation between *expertise* and these two factors, we further compute the rank correlation between them shown in Table 12. We can see that *#tweets* has a very weak correlation with *expertise*, and *#followers* has a moderate correlation with *expertise*. The above indicates that high-expertise celebrities do not necessarily publish more tweets and yet they still exert stronger influence compared to low-expertise celebrities, which indicates that *expertise is more important than participation (or contribution) for social media influence*.

**Table 12** Correlation analysis for *#tweets* v.s. *expertise* and *#followers* v.s. *expertise* by domains.

Measures	Enter.	Scien.	Busi.	Sports
<i>#tweets</i> v.s. <i>expertise</i>	0.11	0.17	0.26	0.22
<i>#followers</i> v.s. <i>expertise</i>	0.49	0.32	0.38	0.40

## 6 Analyzing Across-Domain Influence of Top-Expertise Celebrities

In the above, we have studied within-domain correlations between expertise and social media influence. In this section, we study across-domain correlation analysis. Unlike within-domain influence, it will be meaningless to consider the entire ranking by across-domain influence, since high-expertise celebrities are much more likely to become influential in their non-expertise categories but low-expertise are not as we will show later. Therefore, we focus on the across-domain influence analysis of high-expertise celebrities.

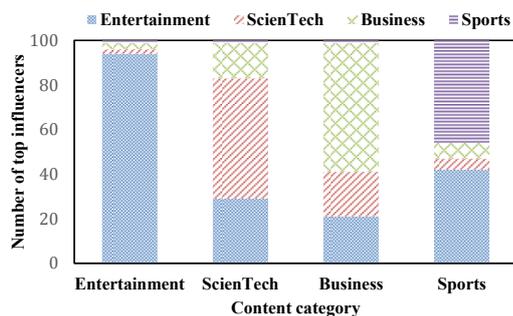
### 6.1 Analysis setup

Given a content category, we rank celebrities from all the four domains by their total influence in a descending order and keep the top 100 celebrities. Note that we only consider those four categories which correspond to celebrities domains for across-domain analysis. We consider a celebrity to be an across-domain influencer if she is ranked in the top 100 most influential celebrities in one of her non-expertise content categories.

We would like to study the overall distribution of celebrities from all the four domains and examine the across-domain influence. Especially, we will focus on high-expertise celebrities in across-domain influencers. Recall that we have split celebrities into 10 bins by their expertise levels in each of four domains, and a high expertise level is indexed by a small number as described in Section 3.1. We consider the celebrities in the first bin of each domain to be high-expertise celebrities and refer to them as *top-10-percent celebrities* for convenience.

### 6.2 The overall distribution

Although we have a total of nine content categories, we only present the results on four content categories which are the same as the celebrities domain labels for easy comparison. The results are shown in Figure 1. We have the following observations: 1) Each content category is still dominated by the celebrities from the domain with the label same as the content category label; 2) There are a considerable number of across-domain influential celebrities in the top 100 positions, indeed some out-of-domain celebrities have even stronger influence than many in-domain celebrities; 3) Celebrities in *Entertainment* tend to be more powerful to influence across domains; 4) Celebrities in one domain are more easily to be influential in other closely related domains. For example, celebrities in *Business* seem to be influential in *ScienTech* while celebrities in *Entertainment* are also influential in *Sports*.



**Fig. 1** The distribution of the top 100 most influential celebrities on four content categories. Columns correspond to content categories while different colors denote celebrity domains.

### 6.3 The proportion of top-10-percent celebrities in across-domain influencers

Having observed that celebrities can have across-domain influence, we further conduct an analysis to find out the proportion of top-10-percent celebrities in across-domain influencers. Intuitively, high-expertise celebrities are more likely to be able to influence across domains than low-expertise celebrities.

For each content category of  $\{Entertainment, ScienTech, Business, Sports\}$ , we first count the total number of across-domain influencers from a given celebrity domain, and then compute the the proportion of celebrities who are from the top-10-percent in that domain. The results are shown in Table 13. We can see the results conform to our intuition well: a large proportion of across-domain influencers are from the top-10-percent celebrities in their own expertise domain, which indicates the strong influence power of the top-10-percent celebrities. To further examine the influence of the top-10-percent celebrities, we compute the proportion of retweet contribution of the top-10-percent celebrities in their own expertise domain: the top-10-percent celebrities nearly received about 70% retweets in their expertise domains.

**Table 13** Proportion of the top-10-percent celebrities who have been ranked in top 100 outside of their expertise domain.

Entertainment	ScienTech	Business	Sports
72.2%	81.3%	87.0%	73.1%

#### 6.4 Why the top-10-percent celebrities are more powerful to influence across domains?

The above analysis shows that top celebrities are potentially influential across domains. However, celebrities with low expertise levels do not appear to have such across domain influence.

First, a traditional explanation of across-domain influence is to refer to “structural holes”. Across-domain influencers are related to “structural holes” defined in sociology [7], who play the role of communicating with information sources and propagating information from multiple domains. One important characteristic of structural holes is they have important links or relationships in different domains. By examining the friend links, we do find that across-domain influencers have more friends in different categories, especially high-expertise celebrities. We present the number of out-domain friends for four domains respectively by expertise levels in Figure 2. We can observe that the first bin has a much larger number of out-domain friends than the other bins for *ScienTech* and *Business*. This observation is consistent with the basic ideas of structural roles, who rely on out-domain links to connect different domains. What is surprising is that for *Entertainment*, the top-10-percent celebrities have relatively few out-domain links. Nevertheless, we still observe the top-10-percent celebrities in this domains have strong across-domain influence even without substantial across-domain friends. It can be partly explained by “status generalization”. Through a process of “status generalization” [28], group members with relatively high social status *outside the group* are expected to be more competent at the group’s task than lower-status group members. In our analysis here, celebrities in *Entertainment* usually has a very large number of followers. As such, any of the tweets of top celebrities in *Entertainment* will receive a significant amount of attention.

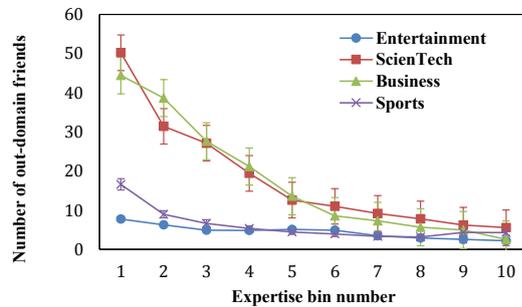


Fig. 2 Number of out-domain friends by four celebrity domains.

In this part, we want to introduce a new perspective to explain why high-expertise are more likely to influence across domains. We try to reveal the differences between high-expertise and low-expertise celebrities by analyzing the propagators of celebrities. Intuitively, if there exist many propagators who have forwarded some tweets

from the expertise category of a celebrity would also forward tweets from her non-expertise categories, then the celebrity is likely to have across domain influence.

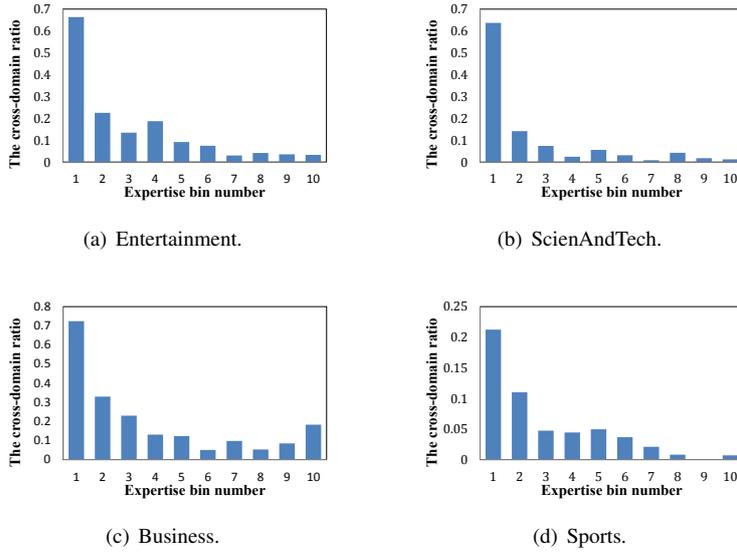
It is not feasible to conduct analysis at the individual level due to data sparsity that most of the common users only forward one or very few tweets of the same celebrity. Therefore, we consider celebrities with the same expertise level in a domain as a whole. As defined in Section 4.1.3, for the celebrities from an expertise domain  $c$ , we use  $\mathcal{F}_{(E)}^c$  to denote the propagators of their tweets classified into the expertise category; while we use  $\mathcal{F}_{(NE)}^c$  to denote the propagators of their tweets classified into non-expertise categories. Now we want to measure how many of  $\mathcal{F}_{(E)}^c$  will further make retweets on celebrities' tweets classified as non-expertise categories. Give the set of expertise propagators  $\mathcal{F}_{(E)}^c$ , we compute the proportion of users in it who have made at least a retweet classified as non-expertise categories, referred to as *across-domain ratio*. Formally, the across-domain ratio of category  $c$  is defined as  $\frac{|\mathcal{F}_{(NE)}^c \cap \mathcal{F}_{(E)}^c|}{|\mathcal{F}_{(E)}^c|}$ , where  $|\mathcal{F}_{(NE)}^c \cap \mathcal{F}_{(E)}^c|$  denotes the total number of across-domain propagators, i.e. the users who have forwarded tweets from both expertise categories and non-expertise categories.

We still consider the 10-bin's split of celebrities according to their expertise levels where bin 1 denotes the high-expertise celebrities. We then calculate the across-domain ratio for celebrities at different expertise levels. We present the results in Figure 3. We can see that propagators of high-expertise celebrities are likely to forward tweets of the same celebrities on their non-expertise topics. Furthermore, across all the four domains, we notice a decrease of the proportion of propagators who forward tweets outside the expertise areas from the celebrities with lower expertise levels. It shows that celebrities with high expertise levels tend to have more "loyal" propagators, who forward tweets of the same celebrities regardless of the actual contents. This phenomenon is referred as to "status generalization" in sociology [28]. These findings explain that why high-expertise celebrities are easier to influence across domains.

## 7 Related work

### 7.1 Influence analysis

Our work lies in the field of influence analysis, which has been a long research direction in sociology [26, 35]. The increasing popularity of online social media makes the evaluation of influence on social media becomes more and more important [11, 22, 25, 20, 29], which is termed as *social media influence* in this paper. Our work is similar to [25] in that we consider topic or categorial influence on online social networks. Our work is also partly inspired by the research on influence measurement on Twitter [9]. The major difference from the previous studies is that we incorporate expertise in influence analysis. It has been previously shown that expertise is an important factor which impacts on task-specific influence, e.g. CQA [38, 34], citation recommendation [13], product promotion [23, 5] and Web search behaviours [30, 31]. These studies usually do not explicitly characterise influence but directly incorporate expertise into the target task. In this way, the expertise itself is related to



**Fig. 3** Across-domain ratio of propagators for four domains in ten different expertise levels.

the studied task, which is less general than the definition of expertise in our work. For the methodologies, they typically extend the existing techniques to incorporate expertise. They construct the test set based on real application data and evaluate the method through the performance of the specific task.

## 7.2 Expertise mining

Another related topic is to mine expertise from social media [12]. Guy et al. [12] made the first analysis of various information signals from social media for expertise and interest mining in an enterprise. Yang et al. [34] propose to joint model topics and expertise in community question answering systems. Kolari et al. [14] utilized the content of the corporate blog posts, their tags, and comments for expertise location. Millen et al. [19] social bookmarking in the enterprise is helpful to increase the awareness of the expertise of other corporate employees. These studies leverage textual semantic information from social media to estimate or learn expertise, which are essentially relevance based methods for expertise location. Our definition of expertise level is closely related to the domain-specific expert status: those team members who are identified as possessing expertise are often afforded power and status [10]. We proposed a domain-specific PageRank method for learning expertise scores of celebrities in a domain and our focus here is the incorporation of expertise for social influence analysis instead of expertise mining.

### 7.3 Interplay between social influence and expertise

In sociology, there is a common belief that expertise is related to social influence or power [10]. There are several studies which try to qualitatively analyse the effects, relationship and roles of these two factors [10, 18, 8, 4, 17]. These studies typically recruit participants and construct validation experiments. Formal null hypothesis were built and then significance test would be employed. The test for participants is made on relatively clear and simple tasks, and the response data for the participants is collected as the input of significance test. Due to high experimental costs, the number of participants is usually around 100. Another limitation of these studies is that strong assumptions have been made to simplify the overall test experiments. There is a lack of validation of such a correlation on large datasets [21] and no prior studies presented a quantitative measurement of the correlation. Without a formal conclusion on the correlation, it is difficult to explain why the involvement of expertise works and how it affects the performance in some tasks.

### 7.4 Connection to social philosophy studies

Our work has close connection to previous studies in social philosophy, including social status [10, 27, 28, 6, 21], expertise effects [32, 24, 16, 27] and structural holes [7]. However, most of these studies were not developed on online social networks and tested on small datasets. Our findings confirm well to these studies and bridge the gap on the study of online social networks by quantitative evidence from real datasets.

In particular, next we discuss the connection between our work and an important theory in social philosophy, i.e., Status Characteristics Theory.

Status Characteristics Theory (SCT) [6, 28] was developed in sociology and social philosophy, which has been considered for better understanding social influence [21]. SCT explains social influence in task groups which is stated as a function of the expectations that group members hold for themselves and for others about their relative ability at the task [21]. Those who are expected to be more competent at the group's task participate more and exert more influence over group decisions than those who are expected to be less competent.

**Connections.** To make a connection to SCT, we can define the “group” to be a category or domain of celebrities, the “task” to be enriching online contents on a category (not necessarily the same as the celebrities' expertise domain) and “competence” to be the number of tweets forwarded by their followers. One important specific status characteristic is the occupation [6]. Thus, in this paper, we consider a celebrity's occupation the same as her expertise area. Our empirical findings conform well to SCT: 1) high-expertise celebrities are expected to be more influential on social media; 2) the influence depends on the relevance between the celebrity's expertise domain and the considered content category: a celebrity will be more influential on their own expertise category; 3) “status generalization” provides a principle explanation of across-domain influencers in sociology [28].

**New observations.** In addition to common points with SCT, our study has shown some new findings in social media. First, status (expertise level) is still *strongly* correlated with celebrities influence even in less relevant (non-expertise) content categories (See Table 7). Second, although high-expertise celebrities are expected to participate more at the group’s task, our results reveal that high-expertise celebrities do not necessarily publish more tweets and yet they still exert stronger influence compared to low-expertise celebrities, specially for average influence (See Table 10). In summary, **status seems to be more important than relevance and participation for social media influence.**

## 8 Conclusions

In this paper, we presented the first study which took into account of expertise into social influence analysis on a large dataset constructed from Sina Weibo, and we proposed a computational approach to measuring the correlation between expertise and social media influence at three levels. We anticipate that this study could provide insights to deeply understand social media influence, and inspire more follow-up research along this line.

Overall, we have found that there is a strong correlation between expertise level and social media influence on both expertise categories and non-expertise categories. Although the overall correlation patterns are similar, the correlation degrees measured in the expertise-related categories are much more significant than those measured on non-expertise categories. Furthermore, top-expertise celebrities are likely to influence across domains and even more influential than the in-domain celebrities. Interestingly, we explain these findings by analyzing the “audience” of celebrities, i.e. the propagators. Our work has close connection with related studies in sociology, and further shows important new findings: status seems to be more important than relevance and participation for social media influence.

Our study is potentially related to many practical applications such as celebrity endorsement. Most business organizations rely on some existing social influence ranking tools such as *Klout* to analyze the expected advertisement effects of given celebrities. Our proposed approach for social media influence provides an alternative way to learn celebrities’ influence. As have been shown in our empirical analysis, our approach provides insights into the classic problem of balancing relevance and reputation in celebrity endorsement [2].

## References

1. Spearman’s rank correlation. [www.mei.org.uk/files/pdf/Spearmanrcc.pdf](http://www.mei.org.uk/files/pdf/Spearmanrcc.pdf) (2007)
2. Angela, B., Maureen, W., Steven, B.: The naked truth of celebrity endorsement. *British Food Journal* **105**, 288–296 (2003)
3. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone’s an influencer: quantifying influence on twitter. In: *WSDM* (2011)
4. Bazarova, N.N., Yuan, Y.C.: Expertise recognition and influence in intercultural groups: Differences between face-to-face and computer-mediated communication. *Journal of Computer-Mediated Communication* **18**(3), 437–453 (2013)

5. Bearden, W.O., Etzel, M.J.: Reference group influence on product and brand purchase decisions. *Journal of Consumer Research* **9**(2), 183–94 (1982)
6. Berger, J., Fisek, M.H., Norman, R.Z., Zelditch, J.M.: *Status Characteristics and Social Interaction: An Expectation-States Approach*. Elsevier (1977)
7. Burt, R.S.: Structural Holes and Good Ideas. *The American Journal of Sociology* **110**, 349–399 (2004)
8. Catizone, R., Guthrie, L., Thomas, A.J., Wilk, Y.: Lie: Leadership, influence and expertise. In: *LREC'12*, pp. 3692–3696 (2012)
9. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: *ICWSM'10* (2010)
10. French, J., Raven, B.: “The bases of social power” in Cartwright. Ph.D. thesis (1959)
11. Goyal, A., Bonchi, F., Lakshmanan, L.V.: Learning influence probabilities in social networks. In: *WSDM '10* (2010)
12. Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., Ronen, I.: Mining expertise and interests from social media. In: *WWW '13* (2013)
13. He, Q., Pei, J., Kifer, D., Mitra, P., Giles, L.: Context-aware citation recommendation. In: *WWW '10* (2010)
14. Kolari, P., Finin, T., Lyons, K., Yesha, Y.: Expert search using internal corporate blogs. In: *Proceedings of the Workshop on Future Challenges in Expertise Retrieval at SIGIR 2008*, pp. 2–5 (2008)
15. Kwak, H., Lee, C., Park, H., Moon, S.B.: What is twitter, a social network or a news media? In: *WWW* (2010)
16. Larson, J.R., Christenson, C., Abbott, A.S., Franz, T.M.: Diagnosing groups: Charting the flow of information in medical decision making teams. *J. Personality Soc. Psych* **71**, 315–330 (1996)
17. Martina, Z., Anne, K.S.: Expertise recognition and influence in intercultural groups: Differences between face-to-face and computer-mediated communication. *HCI in Work and Learning, Life and Leisure* pp. 513–529 (2010)
18. Meagher, L., Lyall, C., Nutley, S.: Flows of knowledge, expertise and influence: a method for assessing policy and practice impacts from social science research. *Research Evaluation* **17**(3), 163–173 (2008)
19. Millen, D.R., Feinberg, J., Kerr, B.: Dogear: Social bookmarking in the enterprise. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06*, pp. 111–120. ACM, New York, NY, USA (2006)
20. Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: *KDD '12* (2012)
21. Oldmeadow, J., Platow, M., Foddy, M., Anderson, D.: Self-categorization, status, and social influence. *Social Psychology Quarterly* **66**(2), 138–152 (2003)
22. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: *WWW* (2011)
23. Senecala, S., Nantela, J.: The influence of online product recommendations on consumers' online choices. *Journal of Retailing* **80**(2), 159–169 (2004)
24. Stewart, D.D., Stasser, G.: Expert role assignment and information sampling during collective recall and decision making. *J. Personality Soc. Psych* **69**(4), 619–628 (1995)
25. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: *KDD'09*, pp. 807–816 (2009)
26. Tedeschi, J.: *The Social Influence Processes*. Aldine treatises in social psychology. Aldine Atherton (1972). URL <https://books.google.com.hk/books?id=9cducD7ZSaYC>
27. Thomas-Hunt, M.C., Ogden, T.Y., Neale, M.A.: Who's really sharing? effects of social and expert status on knowledge exchange within groups. *Management Science* **49**(4), 464–477 (2003)
28. Webster, M., Foschi, M.: *Status generalization : new theory and research / edited by Murray Webster, Jr. and Martha Foschi*. Stanford University Press (1988)
29. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: *WSDM* (2010)
30. White, R.W., Dumais, S., Teevan, J.: How medical expertise influences web search interaction. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, pp. 791–792. ACM, New York, NY, USA (2008). DOI 10.1145/1390334.1390506
31. White, R.W., Dumais, S.T., Teevan, J.: Characterizing the influence of domain expertise on web search behavior. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, pp. 132–141. ACM, New York, NY, USA (2009). DOI 10.1145/1498759.1498819
32. Wittenbaum, G.M.: The bias toward discussing shared information: Why are high-status group members immune? *Communication Research* **27**(3), 379–401 (2000)

33. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on twitter. In: WWW (2011)
34. Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z.: Cqarank: jointly model topics and expertise in community question answering. In: CIKM'13, pp. 99–108 (2013)
35. Zanna, M., Olson, J., Herman, C.: *Social Influence*. Ontario symposium. L. Erlbaum Associates (1987). URL <https://books.google.com.hk/books?id=5uB1NAEACAAJ>
36. Zar, J.H.: Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association* **67**, 578–580 (1972)
37. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: ECIR (2011)
38. Zhou, Y., Cong, G., Cui, B., Jensen, C.S., Yao, J.: Routing questions to the right users in online communities. In: ICDE '09 (2009)