Discovering Dedicators with Topic-Based Semantic Social Networks

Jiyeon Jang[1] **Sung-Hyon Myaeng**[1,2]

[1]Department of Computer Science [2]Division of Web Science and Technology
Korea Advanced Institute of Science and Technology (KAIST), South Korea
{jiyjang, myaeng}@kaist.ac.kr

Abstract

Influential people are known to play a key role in diffusing information in a social network. When measuring influence in a social network, most studies have focused on the use of the graph topology representing a network. As a result, popular or famous people tend to be identified as influencers. While they have a potential to influence people with the network connections by propagating information to their friends or followers, it is not clear whether they can indeed serve as an influencer as expected, especially for specific topic areas. In this paper, we introduce the notion of dedicators, which measures the extent to which a user has dedicated to transmit information in selected topic areas to the people in their egocentric networks. To detect topic-based dedicators, we propose a measure that combines both community-level and individual-level factors, which are related to the volume and the engagement level of their conversations and the degree of focus on specific topics. Having analyzed a Twitter conversation data set, we show that dedicators are not co-related with topology-based influencers; users with high in-degree influence tend to have a low dedication level while top dedicators tend to have richer conversations with others, taking advantage of smaller and manageable social networks.

Introduction

With the growing popularity of online social networking services such as Twitter, Google+ and Facebook, people use them as an important means for sharing information while forming their own online social networks through social interactions and communications. Since online social networks have become important channels for spreading ideas or information as well, they have been regarded as places where users can influence and be influenced by other users. This paper addresses the issue of identifying social network users who are dedicated to communication and interaction over certain topic areas so that directly or indirectly contribute to information diffusion and hence influence others.

Gladwell presented a theory related to a successful spreading of information, referred to as "The Law of the Few", stating that a rapid growth in a certain area usually starts with a handful of people who exhibit some kind of extraordinary behaviors (Gladwell 2002). The theory says there are three types of exceptional people: mavens, connectors, and salesmen. *Mavens* are information specialists, who know everything about a certain topic and love to share what they know. While the mavens are on a leading edge of acquiring and sharing new information, *connectors* are those who know a large number of people to whom new information or discussions on certain issues can be propagated. Lastly, *salesmen* are persuaders who can get people make decisions and take actions. They are strong carriers of infectious ideas, information, and concepts.

Taking an analogy in the context of online social networks, we argue the notions of *experts* and *influencers* that have been studied quite extensively (Kempe, Kleignberg, and Tardos 2003; Kempe, Kleignberg, and Tardos 2005; Zhang, Tang, and Li 2007; Tang et al. 2009; Cha et al. 2010; Weng et al. 2010; Pal and Counts 2011; Bakshy et al. 2011; Purohit et al. 2012) correspond to those of mavens and connectors, respectively. While experts (or mavens) are key sources for new information and ideas, influencers (or connectors) have a great potential to help exposing such information and ideas to others. A variety of methods have been proposed to find the two types of people from online social networks (Tang et al. 2009; Cha et al. 2010; Weng et al. 2010; Pal and Counts 2011; Purohit et al. 2012).

While information diffusion mechanisms and network topology have been the focus of attention in identifying influencers in previous studies, it is not clear whether such influencers indeed participate actively in information diffusion processes. Furthermore, it is difficult to assume that a person simply exposed to a large amount of information can actually promote information diffusion because s/he may suffer from information overload, having to sift

through a large number of messages from the connected people. On the other hand, *salesmen* on online networks would be dedicated to taking an active role of actually getting people engaged in an issue and leading them to take a certain action related to the issue. Despite the importance of identifying salesmen for real influence on other online users, there has been little effort for research along this line.

In this paper, we introduce the notion of *dedicators*, which is similar to that of salesmen in the "The Law of the Few" theory. Our approach is to analyze conversations that have actually been carried out by online social network users, instead of looking at the network topology, because we are interested in observing the actual flow of information, ideas, and issues among the people. Online conversations are important for our purpose because they promote information exchange and social awareness of certain issues. Given that conversations are usually centered around one or more topics and that people have different levels of expertise and influence on different topics (Saez-Trumper 2012 and Tang et al. 2009), it is natural to assume dedicators are topic-dependent.

This paper introduces the notion of the dedicators, which measures the extent to which a user has dedicated to transmit information in selected topic areas to the people in their egocentric networks. Also, we propose a measure that combines both community level and individual level factors, which are related to the volume of and the engagement level with their personal tendency.

Related Work

Numerous studies have addressed the problem of social influence and information diffusion in online networks (Aggarwal 2011). The main focus of the studies has been on the structural properties of the network such as the size and degree of connection distributions (Kempe, Kleinberg, & Tardos 2003; Kempe, Kleinberg, & Tardos 2005; Kumar, Novak, &Tomkins 2006; Kwak et al. 2010; Cha et al. 2010; Tang et al. 2009; Bakshy et al. 2011) primarily to measure influence. While the main stream of the analyses has dealt with syntactic social networks that are based on the existence of explicit connections like following-follower relationships in Twitter, a new line of research has emerged beyond the analysis of the syntactic social networks, focusing on the contents flowing over syntactic networks (Weng et al. 2010; Macskassy 2011; Qi, Aggarwal & Huang 2012; Jang et al. 2012).

Some researchers found that influencers did not dedicate to a specific topic and did not always play a key role in information diffusion (Cha et al. 2010; Bakshy et al. 2011). Bakshy et al. (2011) found that ordinary individuals, who do not have extremely high influence, could play an important to spread information. These findings are in line

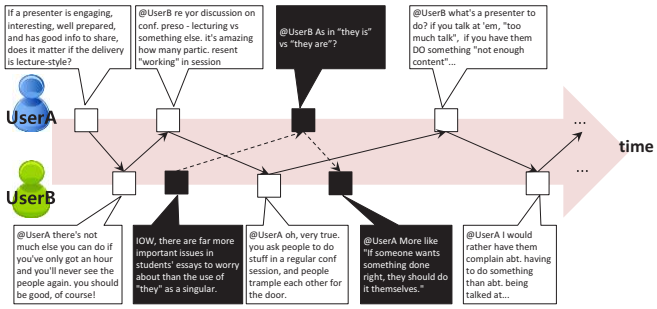

Figure 1. An example of conversations

with our current research.

Although identifying user roles in social networks beyond influencers has been studied (Chan, Hayes, & Daly 2010; Budak, Agrawal, & El Abdadi 2010; Tinati, Carr, & Hall 2012; Saez-Trumper 2012; Purohit et al. 2012), very little work has been done for identifying users serving as salesmen. In a recent work, Budak, Agrawal, & El Abdadi (2010) formally defined the three types of important users (mavens, connectors, and salesmen) based on the "Law of the Few" theory (Gladwell 2002) and added the fourth type referred to as translators who are in charge of making a bridge between different communities or groups. They computed each user role based on the syntactic structure of blog posts with links in them, not the contents.

Our work focuses on identifying the user role of dedicators, which is similar to salesmen. Since the dedicators (or salesmen) should communicate with others, we formally define dedication level with a few factors obtainable from an analysis of conversations. Instead of focusing on the syntactic structure of the networks, we focus on the egocentric semantic network structure.

Data Preparation

Twitter Conversation Dataset

We used a collection of Twitter conversations for our analysis since people use Twitter as a place for exchanging conversations on various topics with other people, groups, and even the public (Java et al. 2007;Honey and Herring 2009; Boyd, Golder, and Lotan 2010; Ritter, Cherry, and Dolan 2010; Chen, Nairn, and Chi 2011). While a conversation can be in various forms depending on the nature of SNS, we define it in Twitter as a thread of sequential tweets connected with the "@" option. An example of conversations in Twitter is shown in Figure 1 where the two threads of white boxes and black boxes are different conversations between two users.

We obtained a sampled user list from the conversation collection constructed by the previous work (Jang et al. 2012). From the original collection consisting of 5,928 users each of which has more than 3,200 tweets in English

Table 1. Dataset description

Total number of users	1,550
Total number of unique conversational partners	224,474
Total number of relationships(dyads)	251,864
Total number of conversations	1,060,981
Total number of exchanged tweets in conversation	6,474,849

in total and at least one conversation between Sep. 9th, 2011 and Oct. 4th, 2011, we selected all the conversations the sampled users were engaged in. The resulting data set contains 4,313,085 conversations for 5,928 users.

To ensure that meaningful topics can be extracted from conversations, we refined our dataset further. We only kept the conversations that contain at least three tweets so that both users replied at least once. Furthermore we filtered out the users having less than 400 conversations, in order to make sure we had enough data for topic extraction using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004; Steyvers and Griffiths 2007). For linguistic processing, we removed stop words and unusual meaningless words such as "aaaaaa", "aaaaaah", or "mossst", and applied the Porter stemmer. In addition, we filtered out the terms whose user frequency is less than ten since some words appeared only in a few users' conversations. As a result, the resulting dataset is relatively cleaned without many spelling errors and idiosyncratic lemmas.

The final dataset we used for this study contains a total of 6,474,879 tweets in 1,060,981 conversations for 1,550 users and 31,820 unique terms. The first conversation in our dataset occurred on Oct. 2nd, 2008, while the last one was on Oct. 4th, 2011. The volume of our dataset is described in Table 1.

Topic-based Semantic Social Network

In order to define and measure topic-dependent dedication levels and ultimately identify dedicators for a topic area, we must first construct egocentric topic-based semantic social networks derived from the conversations as in Jang et al. (2012). Topics are extracted from individual conversations by applying LDA, and each conversation is represented by a topic probability distribution. That is, a conversational relationship between a user pair can be characterized with a topic probability distribution by aggregating the conversations between them. Furthermore, the notions of topic diversity and topic purity were used to enrich each relationship. Topic diversity shows whether a relationship covers a wide range of topics whereas topic purity indicates whether a relationship has a topical focus. The two are useful in characterizing topical interactions between two users. Figure 2 shows a representation of a relationship in the resulting topic-based semantic social networks. The bar shows the topic probability distribution

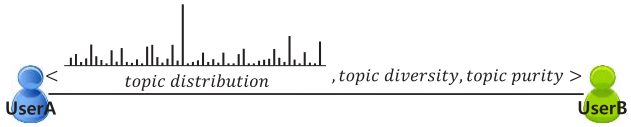

Figure 2. A relationship in a topic-based semantic social network

in the relationship. Along with the topic distribution, the relationship is enriched by topic diversity and topic purity values.

Measuring Dedication Level

Use of Global Topics

After constructing topic-based semantic social networks, each user has a number of relationships with the partners. Each relationship has a probability distribution of 50 topics, and each topic is in turn represented by a probability distribution of words. Since different topic sets have been generated for different egocentric networks independently from each other, however, it is not straightforward to identify topic-based communities for the entire user population from the disparate topical networks.

Instead of applying an ordinary clustering algorithm for 77,500 topics (1,550 users times 50 topics) to identify global topical areas, which is too time-consuming, we devised a novel method for relating all the *local* topics generated for individual networks to a set of common *global* topics generated for the entire set of conversations. We first generate 200 global topics by applying LDA to the set of all the conversations aggregated from the entire user population. The next step is to map each local topic to a subset of the global topics that are sufficiently similar to it.

The Jaccard similarity measure is used to compute similarity between two different word distributions corresponding to local and global topics. In order to minimize the influence of peripheral words in similarity calculations, we select top ranked words for a topic, whose probability values sum up to 0.5. In other words, the probability values of the top ranked words are added sequentially until the sum reaches 0.5 so that the words with sufficiently low probabilities are not considered in computing Jaccard similarity. It turns out that a total of 30 words on average were selected for global topics whereas 82 words on average were used for local topics. A link between local and global topics is established when the Jaccard similarity value is greater than or equal to 0.1. Table 2 illustrates how local topics are mapped to a global topic in our dataset. Note that "kim" and "lil" in local topic #4 and "selena" and "queen" in local topic #5 are singers, and the topics all represent "music".

Table 2. An example of a global topic and its linked local topics from different users from the dataset

Topic	Top-ranked words representing a topic
global topic	{song, listen, music, sing, album, …}
local topic #1	{like, just, song, new, album, …}
local topic #2	{song, love, like, album, listen, …}
local topic #3	{like, listen, music, song, panic, …}
local topic #4	{kim, love, lil, album, plai, …}
local topic #5	{queen, selena, listen, album, think, …}

Linking a local topic to one or more global topics is essential to converting local topic distributions to global ones for a conversational relationship. After this conversion process, all the relationships across different egocentric networks can now have probability distributions over 200 common global topics. For the work of identifying dedicators, however, the level of analysis need to be done for individual conversations rather than relationships resulting from aggregating conversations. In preparation for identifying all the conversations for a particular topic, we eliminate negligible topics whose probabilities are sufficiently low. This is done by applying the same method of selecting top ranked topics (instead of words as in the previous case) whose probability values sum up to 0.5. The average number of topics included for conversations is 2.02.

Now that about two salient topics chosen from the set of 200 global topics have been assigned to each of the conversations in all the egocentric semantic social networks, it is now possible to identify all the users who had at least one conversation for a specific topic. That is, we can build a community of users who have an interest in a particular topic. Put differently, we can analyze conversational behaviors of the users who have exchanged conversations on a particular global topic with a sufficient depth.

Dedication Factors

We now propose to use a set of computable factors that help determine a user's dedication level for a topic. All the factors are based on the egocentric semantic network for a user, in which not only partner relationships but also individual conversations for them are available. After explaining the factors, we will show how they are combined to measure the degree of a user's dedication for one of the 200 global topics.

Volume

In determining a user's dedication level for a topic, the numbers of conversations on the topic and the corresponding conversational partners would play an important role. The more conversations the user was engaged in for a specific topic and the more conversational partners on it, the more dedication s/he had. While the former measures the "depth" of dedication, the latter emphasizes its "breadth". For a topic *i*, we denote the number of conversations and

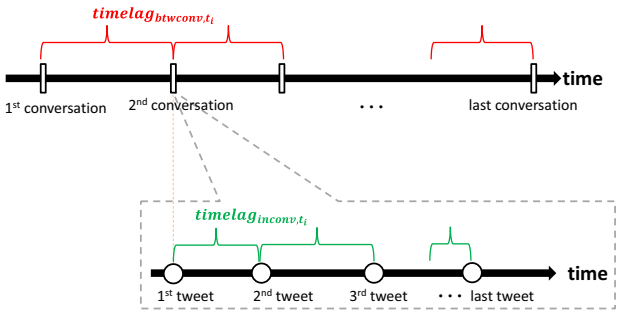

Figure 3. Time lags between conversations and between tweets in a conversation. The gray dashed box shows an example of time lags between tweets in 2nd conversation.

the number of conversational partners as $num_{conv,i}$ and $num_{partner,i}$, respectively. Furthermore, we compute the volume factor for a topic *i* as follows:

$$vol_i = num_{conv,i} * num_{partner,i} \quad (1)$$

Engagement

Another type of factors we consider measures how actively the user was engaged with the partners for conversations on a topic, rather than how long and widely. We consider four factors: the number of conversations per partner ($num_{conv/partner}$), the length of a conversation (len_{conv}), time lag between conversations ($timelag_{btwconv}$), and time lag between tweets in a conversation ($timelag_{inconv}$). The number of conversations per partner indicates the intensity of an interaction. A user having a large number of conversations per partner is likely to have a long-lasting interest in the topic. The length of a conversation, measured by the number of tweets in a conversation, has been used to show how much participants in a conversation are engaged (Chen et al, 2011). A participant having a long conversation is likely to be passionate about the topic.

As a negative sign, we consider time lags between conversations and between tweets in a conversation. A time lag between two conversations is an interval of time between the end of the earlier conversation and the beginning of the next conversation on the same topic, which is shown as the red curly brackets in Figure 3. A time lag between two tweets in a conversation is an interval of time between one tweet and the following tweet in a conversation. It indicates the response time in a conversation, which is shown as the green curly brackets in Figure 3. A short time lag indicates an active user for both cases. In order to compute the time lags between conversations and between tweets for a topic, we take the average of the time lags for all conversations on the topic. Finally, we can compute the engagement factor for a user on a topic *i* as follows:

$$eng_i = \frac{num_{conv/partner,i} * len_{conv,i}}{timelag_{btwconv,i} * timelag_{inconv,i} + 1} \quad (2)$$

Personal Tendency

While the aforementioned factors measure the characteristics of conversational partners, conversations, and tweets, it is important to consider the overall characteristics of a user's conversational behaviors toward a topic. For this aspect, we use *personal tendency* based on topic diversity and topic purity. In the topic-based semantic social networks, each relationship is characterized with topic diversity and topic purity. Topic diversity measures the degree to which a relationship shares a wide range of topics, and topic purity indicates the tendency a relationship focuses on narrow topics (Jang et al. 2012). Users with high topic diversity share a variety of topics with their conversational partners and therefore are likely to lower the dedication level for individual topics. However, users with high topic purity would end up paying high-level dedication to individual topics.

Given a topic, it is easy to identify all the relationships where at least one conversation has it as a salient topic. Since each relationship satisfying the condition has its own topic diversity and purity values, their median values over all the relationships can be taken as a basis for computing the user's personal tendency. For a topic *i*, we compute the personal tendency factor as follows:

$$ten_i = \frac{topic\ purity_i}{topic\ diversity_i} \tag{3}$$

Topic Weight

Another important factor to be considered must come from the probabilities of the salient topics assigned to individual conversations. Given a topic for a user, we can collect all the probability values from the topic representing conversations and compute the topic weight by summing the values and normalizing the result with the number of conversations for the user. This weight can be seen as the level of user interest in the topic and used as an adjustment factor for the others introduced above. For topic *i*, we compute topic weight as follows:

$$w_i = \frac{\sum_{c \in conv_i} p_{c,i}}{num_{conv,i}} \tag{4}$$

where $conv_i$ represents conversations on topic i and $p_{c,i}$ probability of topic *i* in conversation c.

Community vs. Individual Level Dedication

Given that all the factors introduced so far generate a set of values for a user on a topic, a composite measure for dedication to topic *i* can be formulated for a user as follows:

$$D_i^{abs} = w_i \sum_k a_{i,k} \cdot f_{i,k} \tag{5}$$

where $a_{i,k}$ and $f_{i,k}$ ($k \in$ {volume, engagement, personal tendency}) are the importance weights and computed values for factors, respectively. The superscript *abs* means the quantity can be used as an absolute value that can be compared against others computed for other users. Since the values computed for the factors are based on the conversations containing a global topic, it is almost trivial to form a topic-specific community by simply gathering the users who have conversations on that topic. Therefore, the factor values as well as the dedication values computed by (5) are directly comparable among different users in the community.

On the other hand, a single user has multiple absolute dedication levels corresponding to different topics as well as multiple values for individual factors. By comparing these values across different topics, we can determine the extent to which the user made devotions to different topics. Given a user, for example, the number of conversations for topic *i* is relatively small compared to those for other topics although the number can be regarded as very high in the community. This within-user analysis for multiple topics applies to all the other factors introduced above except for the topic weight. For instance, measuring the engagement factor for each of the topics would show the relative dedication levels of the particular kind across the topics. Relative dedication for a topic *i* for a user is measured as follows:

$$D_i^{rel} = w_i \sum_k \frac{a_{i,k} \cdot f_{i,k}}{Z_k} \tag{6}$$

where Z_k is a normalizing factor equivalent to the sum of absolute factor values computed for all the topics, which means $Z_k = \sum_i f_{i,k}$.

The notion of relative dedication levels across different topics is unique and worth considering in identifying topic-dedicators. Even though a user's dedication level for a topic is mediocre at the topic community level, for example, the relative dedication level for the topic is higher than the others for the user. This user is deemed to have a great potential to become a very influential contributor for the topic if she increases her overall activity levels in the social network. On the other hand, using the relative measure alone for dedication analysis would be misleading because a user may be devoted to a particular topic but the volume of conversations, for example, may be very small compared to those in the topic community. Of course, this relative measure would be helpful in analyzing the relative contributions to different topics by the user in an egocentric semantic network.

Finally overall dedication level for a user on a topic *i* is measured as follows:

$$D_i^{final} = \lambda * D_i^{abs} + (1 - \lambda) * D_i^{rel} \tag{7}$$

where λ is the weight of the importance of the community level absolute dedication and set to 0.5.

Analysis of Dedication

Having defined how a user's topic-specific dedication level is measured with several factors computable based on conversations, we examine whether the factors are appropriate to serve the purpose, whether the composite measure in (7) is appropriate in identifying dedicators, and how different it is in comparison with the way influencers are identified. Our analyses were conducted with the Twitter Conversation dataset.

Correlation Analysis

Uniqueness of Factors and Sub-factors

As a way to validate the use of the factors and sub-factors, we examined how they are correlated. The matrix in Table 3 shows pair-wise correlation values between +1 (strong positive correlation) and -1 (strong negative correlation).

We first pay attention to the correlations among the sub-factors for each of the three factors (*volume*, *engagement*, and *personal tendency*). We found that in the case of the volume factor, num_{conv} and $num_{partner}$ have a relatively high correlation (0.708) in the dataset. Nonetheless, we decided to keep both as sub-factors because they have different meanings measuring unique aspects of conversations. While the number of conversations on a topic indicates the depth of interactions, the number of partners represents the breadth. When two users have the same number of conversations with different numbers of partners for a topic, they certainly show different characteristics in conversational behaviors and exhibit different ways they influence others. Conversely, two users having the same number of partners but different conversation counts would be judged to have different levels of dedication and influences to some of the partners. The sub-factors of the other two, *engagement* and *tendency*, were found to have little correlation among them.

The correlation value for $timelag_{btwconv}$ and num_{conv} is negatively high whereas that for num_{conv} and $num_{conv/partner}$ is positively high. The former indicates that the more conversations a user is engaged in, the less time lag between two consecutive conversations. The latter is somewhat expected because of the high correlation between num_{conv} and $num_{partner}$. Nonetheless, the use of $num_{conv/partner}$ and $timelag_{btwconv}$ is still valid because they were introduced to compute the *engagement* factor whereas num_{conv} was used for the *volume* factor.

To ensure that the three factors are unique enough to warrant their use in computing dedication levels, we computed pair-wise correlations among the three. Correlations were computed for individual topics and then averaged. The result in Table 4 shows that there is little correlation in the three way comparisons.

Finally, we compared the adjusting factor *topic weight* against *volume*, *engagement*, and *tendency*. As in Table 5, *topic weight* has no correlation with either *volume* or *engagement.* That is, the quantitative factors, *volume* and *engagement*, are independent of the depth of the user interest estimated with the topic probabilities on the conversations. Even if a user has many conversations on a topic with many partners, they may not necessarily reflect enough depth of user interest. Since *tendency* is computed with topic diversity and topic purity, it is somewhat expected that it has a moderate level of correlation with *topic weight*.

Community vs. Individual Levels

We argued that the absolute dedication values computed for the people in a topic community would be different from the relative ones computed for different topics in an egocentric network at the individual levels. We examined this conjecture by computing correlation between the absolute and relative dedication values with respect to the three factors. When all the users and topics were considered (the first column labeled with "All" in Table 6), the correlation values between the two were all reasonably high for the three factors. That is, it would be difficult to validate the use of dedication at the individual level.

We note, however, that the surface level correlation is misleading because the long tails of low-ranked people for dedication make the correlation values high. The majority

Table 3. Pearson correlation coefficients among sub-factors

	num_{conv}	$num_{partner}$	$num_{conv/partner}$	len_{conv}	$TimeLag_{btwconv}$	$TimeLag_{inconv}$	$TopicDiversity$	$TopicPurity$
num_{conv}	1.000	-	-	-	-	-	-	-
$num_{partner}$	**0.708**	1.000	-	-	-	-	-	-
$num_{conv/partner}$	**0.609**	-0.033	1.000	-	-	-	-	-
len_{conv}	0.172	-0.022	0.290	1.000	-	-	-	-
$TimeLag_{btwconv}$	**-0.668**	-0.569	-0.343	-0.297	1.000	-	-	-
$TimeLag_{inconv}$	-0.122	-0.012	-0.143	-0.158	0.432	1.000	-	-
$TopicDiversity$	0.066	0.029	0.015	0.313	-0.044	0.070	1.000	-
$TopicPurity$	0.423	0.381	0.142	0.210	-0.339	-0.049	-0.140	1.000

p<0.01

Table 4. Pearson correlation coefficients among factors

Factors Compared	Correlation
volume vs. *engagement*	0.182
volume vs. *tendency*	0.143
engagement vs. *tendency*	-0.042

p<0.01

Table 5. Pearson correlation coefficient between each of three factors and a topic weight

Factors Compared	Correlation
volume vs. *topic weight*	-0.006
engagement vs. *topic weight*	0.116
tendency vs. *topic weight*	0.546

p<0.01

of inactive users for a topic in a topic community would have low values for the factors. Likewise, their dedication levels for the topic, i.e. relative values, in terms of the factors would also remain low at the individual level. This type of bias has been recognized in previous social network analyses (Cha et al. 2010).

As a way to avoid such a bias, we computed correlation for the top 10% and 1% users in terms of absolute dedication values. By doing so, it becomes possible to see whether the topic areas for which the users are perceived as top dedicators at the community level are also as important to the users. As in Table 6, the users perceived as top dedicators for a topic in terms of the three factors do not devote themselves to the topic; there must be other topics to which they pay the same or higher level attention. As a result, we believe that it is important to measure the degree of dedication at the individual level so that final dedication values computed with (7) would better predict potential influence of the dedicators.

Characteristics of Dedicators

Comparison between Dedicators and Influencers

The main premise of this research is that there is a distinction between the two notions, dedicators we propose and influencers as defined in the previous research. In order to show the difference and thus the necessity to measure the degree of dedication separately from the conventional notion of influencer, we computed correlation between each of the different versions of dedication measures and a simple but popular method of measuring influence with in-degree for a node in a static network using following and follower links. As in Table 7, there is no correlation between the influencer measure using in-degree and each of the dedication measures. This result confirms that the popularity measure often used as a basis for detecting influencers is completely different from the way dedicators are identified as proposed in this paper.

Table 6. Pearson correlation coefficients for three factors in community vs. individual level (p<0.01)

Correlation	All Users	Top 10%	Top 1%
Volume	0.701	-0.054	-0.050
Engagement	0.653	0.144	-0.104
Tendency	0.499	0.126	0.193

Table 7. Pearson correlation coefficients between influence and different dedication measures (p<0.01)

Influence vs. dedication	Correlation
In-degree vs. D^{final}	-0.030
In-degree vs. D^{abs}	-0.005
In-degree vs. D^{rel}	-0.118

In further analysis, we examined the overlap between top five influencers and top five dedicators. It turns out that only 56 out of 1,550 users across 55 topics were included in the overlap. Even though the 56 users were judged to be top influencers, 18 of them are found to have less than 1,000 followers, while the top five influencers have 49,915 followers on average.

Factors Making Top Dedicators

While the final dedication measure in (7) takes into account six factors, volume, engagement, and tendency at community and individual levels, it would be valuable to understand the extent to which individual factors contribute to making top dedicators. This analysis is also useful in understanding whether dedicators would be determined by particular dominating factors.

When we compared the dedication levels of individual people for a topic and the factors contributing to the people's dedicator ranks, we observed high correlations as in Table 8. Given the high correlations result from the large population of low-ranked people who tend to have low values for the factors, we did the same analysis for top five dedicators chosen for each topic. The result is surprising in that there is no correlation between the final dedication values and the individual factors. It indicates that a high value for a particular factor alone cannot make a dedicator.

In order to have a better understanding about the roles of the factors in making top dedicators, we first collected top five dedicators for each of the 200 global topics. To see how important the factors were in making top dedicators, we obtained statistics about the numbers of the top-5 dedicators and their ranks in terms of the values of each factor. A result showing relative contributions of the factors towards making top dedicator is shown in Figure 4. Each bar represents the total number of top 1 to 5 dedicators in terms of the particular factor. For example, the volume factor at the individual level has the largest number of top-5 dedicators whose volume values are the highest. The segments show the numbers of the first to the fifth top dedicators in terms of the corresponding factors. The relative heights of the bars roughly indicate the degree to which the

Table 8. Pearson correlation coefficient between dedication and each factor. "C" and "I" represent community and individual levels, respectively (p<0.01).

D^{final} vs.	All	Top 5 dedicators
Volume (C)	0.823	-0.006
Engagement (C)	0.630	0.009
Tendency (C)	0.687	0.004
Volume (I)	0.902	0.072
Engagement (I)	0.721	0.121
Tendency (I)	0.700	-0.003

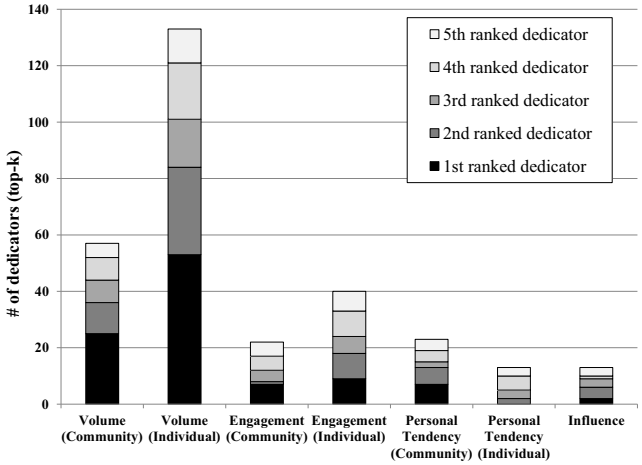

Figure 4. The roles of different factors and the influence measure for identifying top dedicators

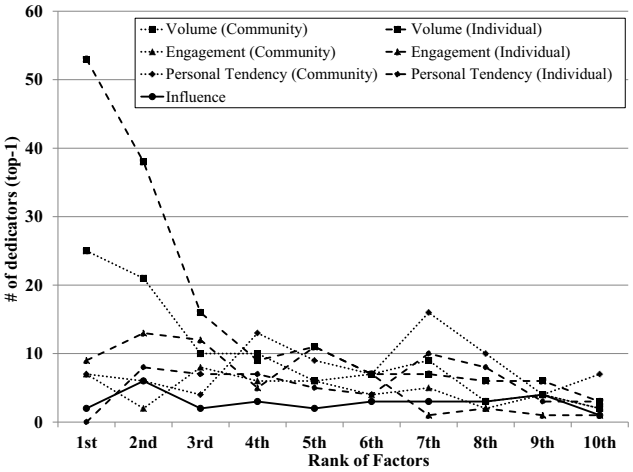

Figure 5. The number of top dedicators supported at different ranks of the six factors and the influence measure

factors contribute towards making top dedicators. Although the volume factors, especially at the individual level, seem dominant, others also make important contributions. The relative values (i.e. individual level) seem to be a more reliable supporter than the absolute values (i.e. community level), which nonetheless is still important to consider in identifying dedicators. Note that the influence measure made the smallest contribution towards making top dedicators.

We conducted a more detailed analysis using the same data so that we can see how the relative rankings of factor values influence top dedicators. As in Figure 5, we plotted the numbers of top dedicators whose factor values are at the top, second, third, and so forth to 10^{th}. For example, the line for *volume (individual)* shows that 53 dedicators had their absolute volume values ranked at the top, and 38 ranked at the second, and so on. The slope going down from the top left to bottom right for the two volumes combined indicates that the volume factors together tend to be a barometer for generating top dedicators. On the other hand, a fair number of top dedicators tend to have the tendency and engagement factor values quote evenly distributed within the top ten.

A final note on this analysis is on the influence measure we compared against the proposed dedication measures. Note that we used in-degree influence measure. The last bar in Figure 4 representing the influence measure is quite short. It indicates that the number of top dedicators who can be selected only based on the influence measure is quite small, perhaps no better than using only the personal tendency at the individual level. The graph in Figure 5 also indicates that the influence measure represented by the red line makes the least contribution toward making dedicators because the top dedicators tend not to have strong values for the influence measure.

Conclusion and Future Work

Users of online social networks play different roles, such as bystanders, active users, influencers, dedicators, trendsetters, etc. In this paper, the focus of our attention has been dedicators for specific topic areas, which are to be compared and contrasted against the conventional notion of influencers. Starting with the theory of "The Law of the Few" that motivated us to identify dedicators in online social networks, we proposed to use a set of factors that help determine dedicators and the final measure that combines absolute and relative degrees of dedication at community and individual levels, respectively.

We conducted detailed analyses with the "Twitter conversation dataset" with which a topical analysis can be done using LDA and egocentric social networks can be built for the users. We first demonstrated that the factors and sub-factors are appropriate to take part in the final measure for determining dedicators and then characterized the dedicators identified with the measure, especially in comparison with the conventional notion of influencers. The analyses all indicate that the proposed dedicator measure is unique and worth exploring further for the purpose of discovering real contributors to spreading information, especially because the measure is derived from topic-based analyses of the conversational content exchanged in online social networks rather than the network structures.

The current study has some limitations. Like the majority of the work on influence analyses, we were not able to verify that the dedicators identified in the automatic method indeed played a key role in diffusing useful information on a topic and/or persuading others to take an action as salesmen would do. A longitudinal study would be required to answer the question. Another limitation is that we

were not able to include the users who would be classified as high influencers in other studies although we used the in-degree measure to identify them. We plan to collect additional data to enable more direct comparisons between dedicators and influencers. We believe that a further study of this kind would reveal more distinct roles of different user groups.

Acknowledgment

This research was supported by a Microsoft Research Asia (MSRA) Faculty-Specific Project and by WCU (World Class University) program under the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

References

Aggarwal, C. C., ed. 2011. Social Network Data Analytics. Springer.

Bakshy, E., Hofman, J.M., Mason, W.A., and Watts, J. D. 2011. Identifying topical authorities in microblogs. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM).

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation, the Journal of Machine Learning Research, 3, pp. 993-1022.

Boyd, D., Golder, S., and Lotan, G. 2010, January. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In 43rd Hawaii International Conference on System Sciences (HICSS).

Budak, C., Agrawal, D., and El Abbadi, A. 2010. Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere. In Proceedings of the First Workshop on Social Media Analytics (SOMA).

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In fourth international aaai conference on weblogs and social media (ICWSM).

Chan, J., Hayes, C., and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. In Proceedings of the fourth International AAAI Conference on Weblogs and Social Media (ICWSM).

Chen, J., Nairn, R., and Chi, E. 2011. Speak Little and Well: Recommending Conversations in Online Social Streams. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI).

Gladwell, M. 2002. The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books.

Griffiths, T. L. and Steyvers, M. 2004. Finding Scientific Topics, Proceedings of the National Academy of Sciences of the United States of America 101, no. Suppl 1, pp. 5228-5235.

Honey, C., and Herring, S. C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In 42nd Hawaii International Conference on System Sciences (HICSS).

Jang, J., Choi, J., Jang, G., and Myaeng, S.H. 2012. Semantic Social Networks Constructed by Topical Aspects of Conversations: An Explorative Study. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM).

Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 work-shop on Web mining and social network analysis.

Kempe, D., Kleinberg, J., and Tardos, E. 2003. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD).

Kempe, D., Kleinberg, J., and Tardos, E. 2005. Influential nodes in a diffusion model for social networks. In Proceedings of the 32nd international Colloquium on Automata, Languages and Programming (ICALP).

Kumar, R., Novak, J., and Tomkins, A. 2006. Structure and Evolution of Online Social Networks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In Proceedings of the 19th international conference on World wide web (WWW).

Macskassy, S. A. 2011. Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis. Social Network Analysis and Mining, 1(4), pp. 355-375.

Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In Proceedings of the fourth ACM international conference on Web search and data mining (WSDM).

Purohit, H., Ajmera, J., Joshi, S., Verma, A., and Sheth, A. 2012. Finding Influential Authors in Brand-Page Communities. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM).

Qi, G. J., Aggarwal, C. C., and Huang T. 2012. Community Detection with Edge Content in Social Media Networks. In IEEE 28th International Conference on Data Engineering (ICDE).

Ritter, A., Cherry, C., and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10).

Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., and Benevenuto, F. 2012. Finding trendsetters in information networks. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).

Sousa, D., Sarmento, L., and Rodrigues, E. M. 2010. Characterization of the Twitter @replies Network: Are User Ties Social or Topical?. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (SMUC '10).

Steyvers, M. and Griffiths, T. L. 2007. Probabilistic Topic Models. T. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds.). Handbook of latent semantic analysis: a road to meaning, Laurence Erlbaum, Mahwah.

Tang, J., Sun, J., Wang, C., and Yang, Z. 2009. Social influence analysis in large-scale networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD).

Tinati, R., Carr, L., Hall, W., and Bentwood, J. 2012. Identifying communicator roles in twitter. In Proceedings of the 21st international conference companion on World Wide Web (WWW), and Mining Social Network Dynamics (MSND) workshop.

Weng, J., Lim, E. P., Jiang, J., and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web Search and Data Mining (WSDM).

Zhang, J., Tang, J., and Li, J. 2007. Expert finding in a social network. In Proceedings of the 12th Database Systems Conference for Advanced Applications (DASFAA).