Modeling Direct and Indirect Influence across Heterogeneous Social Networks

Minkyoung Kim
Research School of Computer Science, The Australian National University
Canberra ACT 0200, Australia
minkyoung.kim@anu.edu.au

David Newth
Commonwealth Scientific and Industrial Research Organisation
Canberra ACT 2600, Australia
david.newth@csiro.au

Peter Christen
Research School of Computer Science, The Australian National University
Canberra ACT 0200, Australia
peter.christen@anu.edu.au

ABSTRACT

Real-world diffusion phenomena are governed by collective behaviors of individuals, and their underlying connections are not limited to single social networks but are extended to globally interconnected heterogeneous social networks. Different levels of heterogeneity of networks in such global diffusion may also reflect different diffusion processes. In this regard, we focus on uncovering mechanisms of information diffusion across different types of social networks by considering hidden interaction patterns between them. For this study, we propose dual representations of heterogeneous social networks in terms of *direct* and *indirect* influence at a macro level. Accordingly, we propose two macro-level diffusion models by extending the Bass model with a probabilistic approach. By conducting experiments on both synthetic and real datasets, we show the feasibility of the proposed models. We find that real-world news diffusion in social media can be better explained by direct than indirect diffusion between different types of social media, such as News, social networking sites (SNS), and Blog media. In addition, we investigate different diffusion patterns across topics. The topics of Politics and Disasters tend to exhibit concurrent and synchronous diffusion by direct influence across social media, leading to high relative entropy of diverse media participation. The Arts and Sports topics show strong interactions within homogeneous networks, while interactions with other social networks are unbalanced and relatively weak, which likely drives lower relative entropy. We expect that the proposed models can provide a way of interpreting strength, directionality, and direct/indirectness of influence between heterogeneous social networks at a macro level.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

General Terms

Theory, Experimentation

Keywords

Macro-level diffusion models, direct/indirect influence, heterogeneous social networks, social media

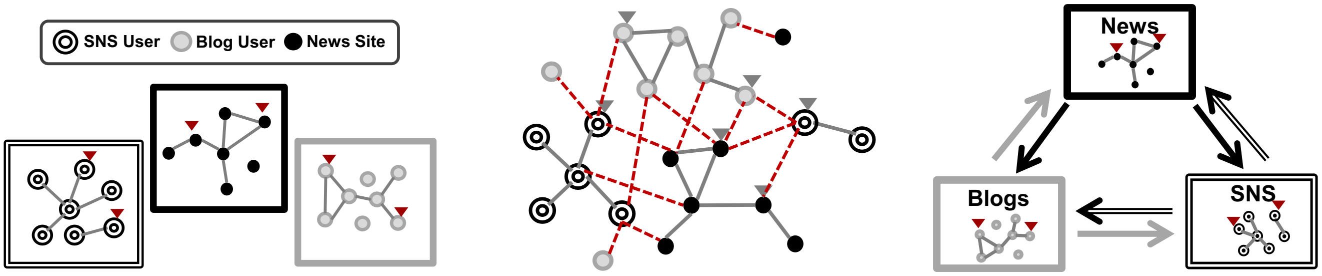
1. INTRODUCTION

Increasingly, web documents are both locally and globally interconnected by sharing information across multiple online social networks. The underlying connections can be defined with diverse user behaviors such as hyperlinks [8, 12, 15], shared quotes [14], similar keywords [3, 10], and specific actions such as retweets and hashtags [18, 21]. Such collective behaviors consequently form a far-reaching diffusion space which possibly ranges from single social networking platforms such as Twitter and Facebook [18, 21, 22] or the blogosphere [3, 10, 15] to multiple different kinds of social networks [8, 12, 14].

In this study, we focus on hyperlinks or written URLs in the main text of a web document as explicit spreading behaviors. Also, we consider social media as a diffusion space, not limited to single social platforms, due to its coverage of heterogeneous social networks such as closely connected News media, friend networks within social networking sites (SNS), and the blogosphere in Blog media. This enables to reveal underlying mechanisms of global diffusion across multiple social networks, but with following main challenges. (1) Underlying structures of multiple social networks are hidden. (2) Diversity of social networks requires meta-population schemes, which enables to classify diverse social networks and interpret macro-level diffusion patterns [4]. (3) Diffusion processes would be varied by topics of information [18, 21].

In order to resolve the issues, we propose dual representations of social networks by defining hidden interaction patterns between different networks in terms of *direct* and *indirect influence*, as shown in Fig. 1. Such interactions have been increasing nowadays with the help of Web technologies such as RSS news feeds, social media aggregators (e.g., TweetDeck and HootSuite), and miscellaneous mobile applications, which enable online users to save the efforts jumping from one social networks to another. Based on the two conceptual frameworks, we model macro-level diffusion with a probabilistic approach by combining structural connectivity and heterogeneity of social networks, which deals with the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
The 7th SNA-KDD Workshop '13 (SNA-KDD'13), August 11, 2013, Chicago, United States.
Copyright 2013 ACM 978-1-4503-2330-7 ...$15.00.



(a) Isolated diverse social networks (b) Direct influence (c) Indirect influence

Figure 1: Dual representations of heterogeneous networks in social media; (a) shows isolated homogeneous social networks on the Web, (b) represents direct interactions (dashed lines) between different types of individuals as if they were in the same networks in a wider diffusion space than their original social networks (thus, external influences [red/dark triangles in (a)] are now considered as internal influences [gray/light triangles in (b)] in globally interconnected social networks), and (c) keeps current network structures, each of which as a whole network indirectly influences other social networks through the increasing rate of adopting behaviors leading to more frequent exposures of information to other networks.

first two challenges. Details are explained in Section 3 and 4. Regarding the third issue of different diffusion patterns by information topics, we analyze topical variations in news diffusion across social media. For comprehensive context of diffusion, we identify real-world news using the Wikipedia Current Events [2] which covers representative topics of conventional news outlets.

Our observation and analysis are based on the ICWSM'11 Spinn3r dataset [1] which contains over 386 million Web documents covering a one-month period in early 2011. In order to understand macro-level diffusion patterns across diverse social networks, we classify social media into News, SNS, and Blog media as representative media types based on its definition and classification by [11]. The authors of [11] defined the term of social media as "a group of internet-based applications allowing the creation and exchange of user generated contents on the ideological and technological foundations of Web 2.0", which helps to avoid misunderstanding the concept of social media.

From experiments on both synthetic and real datasets, we find that news diffusion in social media is attributed to globally and directly interconnected online social networks. By quantifying time-evolving heterogeneity of networks in information diffusion, we interpret dynamics of influence in accordance with different levels of heterogeneity. That is, topics of Politics and Disasters tend to drive concurrent and synchronous diffusion across different social networks by balanced interactions among them, leading to high relative entropy of diverse network participation. On the other hand, Arts and Sports related topics come up with strong internal communications within homogeneous networks but unbalanced and weak interactions with other social networks, likely driving lower relative entropy of disproportionate media participation. Such observations, to the best of our knowledge, are seen for the first time. We expect that the proposed models can apply to a wider class of diffusion phenomena and provide a way of interpreting the dynamics of meta-populations in terms of strength, directionality, and direct/indirectness of influence at a macro level.

In the rest of the paper, Section 2 reviews related work. Section 3 proposes dual representations of heterogeneous social networks, and accordingly Section 4 models two macro-level diffusion mechanisms. Section 5 describes data preparation and measures evolving heterogeneity of social networks in global news diffusion. Section 6 presents experimental results and discusses about outcomes, and finally Section 7 concludes this study.

2. RELATED WORK

The diffusion process of information has been commonly believed to consist of mainly two phases as emergence of information by *external influence* (e.g., mass media) and cascading spreads of the information through *internal influence* (e.g., interpersonal communication) [6, 13, 14, 18, 16]. In the marketing literature this idea was mathematically represented with a conditional likelihood of adoption by the Bass model [5] which consists of likelihoods of "innovation" and "imitation" that correspond to the external and internal influence, respectively. The Bass model has been one of the most influential diffusion models due in part to realistic and robust estimation of new product growth patterns [6]. Its fundamental assumption is that a population is homogeneous and fully connected in the same way as the traditional macro-level diffusion models [16, 19]. This simplicity has enabled intuitive interpretation and has led to a wide range of extensions of the model.

One extension allows heterogeneous mixing of populations such as multinational diffusion of a product. For instance, the adoption rate of a consumer product in one country indirectly influences that in another country [13, 20]. However, this extension disregards the effect of network topologies on diffusion. The other extension is to incorporate network properties such as degree distributions into the stochastic nature of the diffusion process within a contact network [16], but it is still limited to single social networks. This study models macro-level diffusion by combining heterogeneity and structural connectivity of social networks with a probabilistic approach. Also, we generalize the Bass model into dynamics of meta-populations reflecting both direct and indirect influences across multiple social networks.

Diffusion patterns have significant variations across topics [8, 18, 21]. According to [18], political topics are considerably driven by the external influence, while other topics such as entertainment are driven by internal communications. [21] showed that political topics are persistent relative to non-controversial subjects. However, these studies have focused on single social networks, so the dynamics of external and internal influences are limited to the local observations. In this context, our study examines macro-level diffusion patterns of news contents in accordance with six topics, i.e. Politics, Business & Economy, Disasters, Arts & Culture, Sports, and Technology.

3. DUAL REPRESENTATIONS OF HETEROGENEOUS SOCIAL NETWORKS

More and more users come across news stories from diverse social media with the help of the Web technologies such as RSS news feeds, social media aggregators (e.g., TweetDeck, HootSuite) and miscellaneous mobile applications without the need to jump from one to another. Such real-world environments make online social networks increasingly connected with one another. Accordingly, hidden interaction patterns between different social networks are defined in two different ways as *Direct* and *Indirect Influence*.

3.1 Dichotomous View

From the aspect of a single social platform, the world is divided into inside and outside of each platform, and thus it does not distinguish the types of social networks outside, as shown in Fig. 1(a). However, direct and indirect influence between different social networks is not negligible especially when it comes to global diffusion. For a better understanding of underlying mechanisms of diffusion, it would be worth to consider hidden interactions between heterogeneous social networks with a bird's-eye view, away from this dichotomous view.

3.2 Direct Influence

We define a framework of "Direct Influence" in which different types of social networks directly interact with each other as if they were in the same networks, as shown in Fig. 1(b). Due to the collapse of diffusion boundaries of single social platforms, external influences in original social platforms (red/dark triangles in Fig. 1(a)) turn into internal influences between different types of individuals (gray/light triangles in Fig. 1(b)) by their hidden interactions (dashed lines in Fig. 1(b)), and instead new external influences are introduced from outside of the connected different types of networks. This framework interprets influence between different types of social networks as direct and simultaneous effects on diffusion.

External influence has been regarded as a fixed constant, but recently [18] quantified the exogenous out-of-the-network effects. Interestingly, it was shown that almost 30% of information volume in Twitter is attributed to external factors. This figure is ten times larger than the typical value of external influence (0.03) and is rather similar to the average value of internal influence (0.38) in the marketing literature [17]. Such a large proportion of out-of-the-network effects supports the fact that the influence outside of single social platforms is not only negligible but also direct and synchronous.

3.3 Indirect Influence

Fig. 1(c) illustrates the second framework of "Indirect Influence". It keeps the existing social structures as shown in Fig. 1(a), but instead it distinguishes the types of social networks outside of their own networks different from the dichotomous views of single social platforms. In addition, different types of individuals do not directly interact with each other unlike in the case of direct influence, but each social network as a whole indirectly influences other social networks with different levels of strength. This framework interprets influence from different kinds of social networks as indirect and asynchronous effects on diffusion.

In the marketing literature sales growth of a new consumer product in one country is likely influenced by popular diffusion of the product in neighboring or technology-leading countries [13] with different levels of influence without direct interactions between consumers in different countries. This case reflects multinational diffusion driven by indirect influence. Analogous cases can be thought of in news diffusion across social media; different types of social networks are considered as different countries, and the rapid growth of diffusion of some news content in the blogosphere has an effect on increasing diffusion rate of the news in SNS.

4. PROPOSED MODELS

In this section we propose two macro-level diffusion models which reflect direct and indirect influence discussed in the previous section. We first describe a fundamental framework as a background and then extend it to Direct and Indirect Influence Models.

4.1 Fundamental Framework: Bass Model

Let $A(t)$ be the number of cumulative adopters at time t, $a(t) = \mathrm{d}A(t)/\mathrm{d}t$ be the number of new adopters, and n be the whole population. Accordingly, we denote the proportion of the cumulative adopters by $F(t) = A(t)/n$ and the proportion of new adopters by $f(t) = \mathrm{d}F(t)/\mathrm{d}t = a(t)/n$. Then, the ratio of new adopters to potential adopters at time t is called the hazard function $h(t)$,

$$h(t) = \frac{a(t)}{n - A(t)} = \frac{f(t)}{1 - F(t)} \ . \qquad (1)$$

The Bass model [5, 6] assumes the hazard function to be a linear form of the proportion of the cumulative adopters,

$$\frac{f(t)}{1 - F(t)} = p + qF(t) \ . \qquad (2)$$

The parameter p is called the *coefficient of innovation* since it corresponds to the constant proportion of innovators in potential adopters, and q is called the *coefficient of imitation* because it represents the influence of previous adopters [6].

In contrast to the strong assumption of the Bass model that every individual is equally influenced by the population who has previously adopted, the authors of [16] applied the network property of degree distribution to the model to reflect a realistic circumstance in which internal influence varies with an individual's contact network. However, real-world diffusion is hard to define with single social networks alone since dynamic interactions between heterogeneous social networks are not negligible. For a better understanding of diffusion, we propose two macro-level diffusion models based on the concepts in Fig. 1 by combining heterogeneity and structural connectivity of social networks with a probabilistic approach, which improves the accuracy of diffusion mechanisms.

4.2 Direct Influence Model

For modeling diffusion in heterogeneous social networks, we begin with interpreting the Bass model in a probabilistic point of view. Since the proportion of adopters in the Bass model is in fact its expectation in the mean-field mass-action kinetics of the model, it can be thought of as an adoption probability that an average individual adopts at time t,

$$F(t) = P(adopt \mid t), \tag{3}$$

where *adopt* is a binary random variable for the event of an individual's adoption, and it will be abbreviated to "a" in the rest of the paper for brevity. Similarly, we can view the hazard function as a new adoption probability $P(a \mid \neg a, t)$,

$$\frac{f(t)}{1 - F(t)} = \frac{\partial_t P(a \mid t)}{1 - P(a \mid t)} = P(a \mid \neg a, t), \tag{4}$$

where ∂_t denotes the partial derivative with respect to t, and $\neg$ stands for the opposite. Therefore, $P(a \mid \neg a, t)$ indicates the probability that an average individual, who has not adopted before, adopts at time t.

By separating external and internal influences and applying the probability of union of two independent events ($P(A \cup B) = P(A) + P(\neg A)P(B)$), we get

$$\frac{\partial_t P(a|t)}{1 - P(a|t)} = \begin{array}{l} P_{\text{ext}}(a|\neg a, t) + \\ (1 - P_{\text{ext}}(a|\neg a, t))P_{\text{int}}(a|\neg a, t), \end{array} \tag{5}$$

where P_{ext} and P_{int} denote the new adoption probabilities by external and internal influences, respectively.

To deal with heterogeneity of populations, we introduce a random variable, $i = 1, ..., m$ for different types of m meta-populations, and thus construct m different equations of new adoption probabilities for each type as

$$\frac{\partial_t P(a|i, t)}{1 - P(a|i, t)} = \begin{array}{l} P_{\text{ext}}(a|\neg a, i, t) + \\ (1 - P_{\text{ext}}(a|\neg a, i, t))P_{\text{int}}(a|\neg a, i, t) \end{array}. \tag{6}$$

Like the coefficient of innovation in the Bass model, we consider the new adoption probability by external influence as

$$P_{\text{ext}}(a|\neg a, i, t) = p_i, \tag{7}$$

where $p_i \in [0, 1]$. Now, let us focus on internal new adoption probability by considering the structural connectivity of contact networks. Suppose that an individual of type i has k neighbors in which $\mathbf{j} = (j_1, ..., j_m)^{\mathrm{T}}$ neighbors of each individual type have already adopted. Then, from the sum rule and Bayes' theorem, the internal new adoption probability is factorized by

$$\begin{aligned} P_{\text{int}}(a|\neg a, i, t) &= \sum_{k=1}^{n-1} \sum_{\mathbf{j}} P(a, \mathbf{j}, k|\neg a, i, t) \\ &= \sum_{k=1}^{n-1} \sum_{\mathbf{j}} P(a|\mathbf{j}, k, \neg a, i, t) P(\mathbf{j}|k, \neg a, i, t) P(k|\neg a, i, t), \end{aligned} \tag{8}$$

where $n = \sum_{i=1}^{m} n_i$ and n_i is the population of type i.

The distribution of an individual's exposures to previous adopters in its neighbors is modeled as a binomial distribution, which is consistent with prior diffusion models [16, 18]. Thus, each contagion is a Bernoulli trial, and the probability that an individual adopts after $\mathbf{j} = (j_1, ..., j_m)^{\mathrm{T}}$ contacts is

$$P(a|\mathbf{j}, k, \neg a, i, t) = 1 - \prod_{i'=1}^{m} (1 - c_{i'i})^{j_{i'}}, \tag{9}$$

where $c_{i'i} \in [0, 1]$ denotes the probability that an individual of type i adopts when it is exposed to a previous adopter of type i'. Note that it is the probability that an individual is affected by at least one of its adopting neighbors, i.e. one minus the probability of the complementary event that it is not affected by any of the previous adopters in its neighbors.

From a macro point of view, the probability distribution of having $\mathbf{j}$ adopters in k neighbors is a multinomial distribution,

$$P(\mathbf{j}|k, \neg a, i, t) = \frac{k!}{j_1! \cdots j_m! (k - j)!} \prod_{i=1}^{m} P(a|i, t)^{j_i} (1 - P)^{k-j}, \tag{10}$$

where $j = \sum_{i=1}^{m} j_i$ and $P = \sum_{i=1}^{m} P(a|i, t)$.

Finally, we assume that the degree distribution of an individual follows a power law since real-world networks are scale-free networks exhibiting power-law distributions [9, 19]. From our previous study [12], we also found that the distribution of hyperlink cascades across social media follows a power-law.

$$P(k|\neg a, i, t) = \frac{1}{\zeta(\alpha_i)} k^{-\alpha_i}, \tag{11}$$

where α_i is the power law coefficient of individual type i, and $\zeta(\alpha_i) \equiv \sum_{k=1}^{n-1} k^{-\alpha_i}$. Substituting Equation (9), (10), and (11) into Equation (8) gives the internal new adoption probability,

$$P_{\text{int}} = 1 - \frac{1}{\zeta(\alpha_i)} \sum_{k=1}^{n-1} \frac{\left(1 - \sum_{i'=1}^{m} c_{i'i} P(a|i', t)\right)^k}{k^{\alpha_i}}, \tag{12}$$

where P_{int} is short for $P_{\text{int}}(a \mid \neg a, i, t)$. Again, by substituting Equation (7) and (12) into Equation (6), we obtain the system of partial derivative equations for the Direct Influence Model. It is not mathematically tractable, and thus we need to solve it numerically to get the adoption probabilities $\{P(a \mid i, t)\}_{i=1}^{m}$.

4.3 Indirect Influence Model

Now, we turn our attentions into the second conceptual model in Fig. 1(c), where information spreads within homogeneous networks through the connections between users, and its popularity in one social network affects its diffusion in other types of social networks indirectly. Let us focus first on the intra-network diffusion and then consider the inter-network diffusion later.

We begin with the general diffusion model of heterogeneous networks in Equation (6) and make the same assumption that the external new adopter probability is constant as Equation (7) in the Direct Influence Model.

For the internal new adoption probability, we follow similar arguments of the Direct Influence Model except that now individuals are directly connected within homogeneous networks. Suppose that in a network of type i, an individual has k neighbors in which j neighbors have already adopted, and they all belong to the same network of type i. Then, the internal new adoption probability is factorized by

$$\begin{aligned} P_{\text{int}}(a|\neg a, i, t) &= \sum_{k=1}^{n_i - 1} \sum_{j=1}^{k} P(a, j, k|\neg a, i, t) \\ &= \sum_{k=1}^{n_i - 1} \sum_{j=1}^{k} P(a|j, k, \neg a, i, t) P(j|k \neg a, i, t) P(k|\neg a, i, t). \end{aligned} \tag{13}$$

Table 1: Identified real-world news corresponding to our dataset period; news contents and categories are based on the Wikipedia Current Events [2].

Category	Real-world News Stories (January 2011)
Politics	Protests in Tunisia, Egypt, Sudan, and Yemen; Conflicts between Muslim and Christian in Egypt; Tucson shooting; A suicide bombing at Domodedovo International Airport in Moscow; Anti-government activities, Julian Assange Wikileaks
Business and Economy	Unemployment; Food and oil price rising and crisis; US bank crisis; US-China export deal; Energy consumption
Disasters	Floods in Australia, Sri Lanka, and Brazil; Massive winter storm in US; Haiti earthquake; Global warming
Arts and Culture	Academy Movie Awards; Golden Globe Awards; Screen Actors Guild Awards; MSNBC's contract termination with their cable news host; Multiculturalism fail; Asian education
Sports	NFL playoffs; BCS Championships; Australian Open; Sky Sports sexism scandal
Technology	Google technology news; Apple iPad release; iPad for education

Recognize that unlike the Direct Influence Model, where the previous adopters in neighbors are classified by their individual types as $\mathbf{j} = (j_1, ..., j_m)^{\mathrm{T}}$ in Equation (8), the previous adopters in neighbors are of the same individual type, thus collapsed into j in Equation (13).

The Bernoulli influence model of previous adopters in neighbors is the same in the Indirect Influence Model. However, we only consider influence of the same type of users,

$$P(a|j, k, \neg a, i, t) = 1 - (1 - c_i)^j, \qquad (14)$$

where $c_i \in [0, 1]$ denotes the probability that an individual of type i adopts when he or she is exposed to a previous adopter.

The homogeneity converts the probability distribution of adopters in neighbors from a multinomial distribution to a binomial distribution,

$$P(j|k, \neg a, i, t) = \frac{k!}{j!(k-j)!} P(a|i, t)^j (1 - P(a|i, t))^{k-j} \ . \qquad (15)$$

The assumption on the degree distribution of an individual who has not adopted at time t is identical to Equation (11) of the Direct Influence Model. Substituting Equation (14), (15), and (11) into Equation (13) gives the internal new adoption probability,

$$P_{\text{int}}(a \mid \neg a, i, t) = 1 - \frac{1}{\zeta(\alpha_i)} \sum_{k=1}^{n_i - 1} \frac{(1 - c_i P(a|i, t))^k}{k^{\alpha_i}} \ . \qquad (16)$$

Again, by substituting Equation (7) and (16) into Equation (6), we obtain the system of partial derivative equations for the intra-network diffusion of the Indirect Influence Model.

Finally, for the inter-network diffusion, we model the influence of information popularity in one network to diffusion in other networks with a multiplying factor,

$$\frac{\partial_t P(a|i, t)}{1 - P(a|i, t)} = P(a|\neg a, i, t) \left(1 + \sum_{\substack{i'=1 \\ i' \neq i}}^{m} b_{i'i} \partial_t P(a|i', t) \right), \qquad (17)$$

where $b_{i'i}$ denotes the impact coefficient of the network of type i' on the diffusion of the network of type i.

Multiplying a probability by a factor is certainly not a probabilistic way. However, it is a well-known method for extending the Bass model such as pricing and advertising effects [7] and multinational diffusion [13]. To make the new adoption probability meaningful, we impose a probability constraint, $0 \leq \partial_t P(a|i, t)/(1 - P(a|i, t)) \leq 1$.

The system of partial derivative equations for the Indirect Influence Model in Equation (17) is also mathematically intractable. So, we solve it numerically to get the adoption probabilities $\{P(a \mid i, t)\}_{i=1}^{m}$.

In this section, we modeled macro-level diffusion based on the dual representations of heterogeneous social networks. This is a generalization of the simple mass-action Bass model into dynamics of meta-populations in a probabilistic way by combining the two essential features, heterogeneity and structural connectivity of social networks.

5. NEWS DIFFUSION IN SOCIAL MEDIA

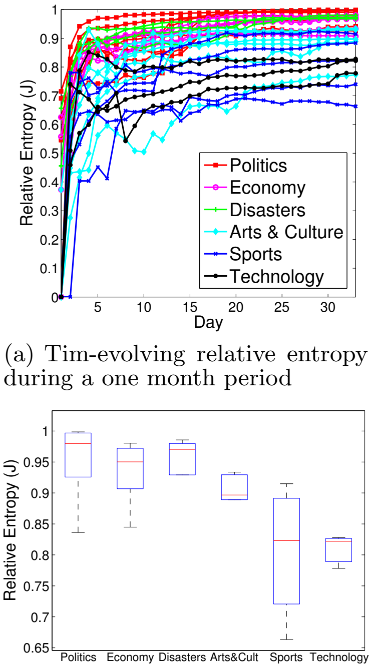
In order to understand underlying mechanisms of diffusion across different social networks at a macro level, we take real-world cases from global spreads of news in social media. In this section, we explain data preparation with its fundamental statistics, quantify evolving heterogeneity of social networks in global diffusion using the measure of *relative entropy*, and finally examine how the levels of heterogeneity are varied by news topics.

5.1 Data Preparation

Our analysis and observations are based on the ICWSM'11 dataset [1] which is freely available to research communities and was collected by Spinn3r, a licensed social media crawler. This dataset consists of over 386 million web documents covering a one-month period in early 2011 (13th Jan to 14th Feb). Each collected document includes a timestamp, language, and a HTML body with hyperlinks.

In our previous study [12], we constructed document networks (4.1 million) by extracting hyperlinks or written URLs in the main text of each document due to large portions of spam hyperlinks even in a single web page. We also labeled destination documents of hyperlinks with noteworthy news by using the Wikipedia Current Events [2] which provides chronologically organized event profiles. These processes allow us to trace meaningful diffusion with much less noise.

We focus on analyzing News, SNS, and Blog documents (98.37% of the original dataset) because these are not only the most relevant to real-world news, but we can also ob-



(a) Tim-evolving relative entropy during a one month period

(b) Distribution of relative entropy at the final time step (t=33)

Figure 2: Time-evolving heterogeneity of diverse social networks (News, SNS, and Blog types) involved in news diffusion; each news content in (a) is color-coded by six categories in Table 1.

serve dynamic interactions among representative types of social media. From the labeled documents, we obtained the largest connected document network (650K nodes) and its corresponding user networks (94K nodes) as well. From the largest document network, we selected the top 35 news stories which have driven adoptions of at least 300 identified users. Table 1 shows the selected news contents with their categories based on the Wikipedia event profile.

5.2 Measuring Evolving Heterogeneity of Social Networks in News Diffusion

We examine the degree of heterogeneity of social network participation in diffusion of news stories, and also observe its variations by topics. By using the *relative entropy* as a measure of dispersion [23], we quantified time-evolving heterogeneity as Equation (18).

$$J(t) = \frac{H(t)}{H_{max}} = \frac{1}{H_{max}} \left(- \sum_{i=1}^{m} p_i(t) \log_2 (p_i(t)) \right), \qquad (18)$$

where the entropy $H(t)$ is a measure of dispersion at time t, and $p_i(t)$ is the proportion of cumulative users of media type i among m types at time t. When the proportions of each media type users at time t are uniformly distributed, the entropy is highest ($H_{max} = log_2 m$), and it is lowest ($H_{min} = 0$) when only one type of media exists. Therefore, we can quantify time-evolving heterogeneity of diverse social network participation in diffusion by relatively comparing the entropy $H(t)$ at time t with the maximum entropy H_{max}.

As shown in Fig. 2(a), the heterogeneity (J) generally increases as time evolves, which means that different media type users eventually participate in news diffusion in a balanced way. This increasing heterogeneity is possibly led by direct or indirect influence between different social networks, which supports the importance of modeling diffusion across heterogeneous networks.

Table 2: Averages (standard deviations in parenthesis) of model fitting errors (RMSE) of each diffusion model with two groups of synthetic datasets (50 for each), where Group 1 is generated by DM and Group 2 by IM (BM: Bass Model, DM: Direct Influence Model, IM: Indirect Influence Model); bold fonts represent the best performance in each group.

	BM	DM	IM
Group 1	5.437e-3 (2.477e-4)	**2.796e-3** **(2.115e-4)**	3.113e-3 (2.319e-4)
Group 2	4.671e-3 (1.173e-3)	2.877e-3 (1.167e-3)	**2.729e-3** **(9.842e-4)**

However, there exist variations by topics, as shown in Fig. 2(b). News contents such as Politics, Economy, and Disasters drive higher relative entropy (balanced proportions of different media types), while the topics such as Technology, Arts, and Sports lead to comparatively disproportionate distributions, i.e. lower relative entropy.

6. EXPERIMENTS AND DISCUSSION

We evaluate our proposed models using both synthetic and real datasets and compare the results with the Bass model as a baseline. We fit the models by minimizing the sum of squared errors in an iterative way until the error converges. As evaluation metrics, model fitting errors and parameter errors are used. After verification of parameter recovery with synthetic datasets, we interpret news diffusion patterns in social media with inferred parameter values from real datasets.

6.1 Experiments on Synthetic Data

In the previous section, we introduced the Bass Model (BM) and proposed two macro-level diffusion models, Direct Influence Model (DM) and Indirect Influence Model (IM). Now, the objective of this section is to recover the hidden diffusion patterns, i.e. how different types of individuals have interacted with one another *directly* or *indirectly*, when the cumulative number of adopters at every time step is given as an input.

For that purpose, we generated two groups of synthetic data sets based on the proposed diffusion models and fitted them to each group of the datasets. From the observation of real datasets, we selected 50 parameter sets for each model (100 datasets in total), and generated the numbers of adopters, $\{(A_n(t), A_s(t), A_b(t))\}_{t=1}^{T=30}$ for three media types, News, SNS and Blog, respectively. The length of time step T is chosen as one month (30 days) in order to reflect our real dataset period.

Let us denote the model parameters by $\boldsymbol{\theta}_i (i = n, s, b)$, where their definitions are different in each diffusion model; $\boldsymbol{\theta}_n = \{p_n, q_n\}$ in the BM, $\boldsymbol{\theta}_n = \{p_n, c_{nn}, c_{sn}, c_{bn}\}$ in the DM, and $\boldsymbol{\theta}_n = \{p_n, c_n, b_{sn}, b_{bn}\}$ in the IM. In order to fit each model to datasets, we apply the Nonlinear Least Squares (NLS) which minimizes the normalized root mean squared errors (RMSE),

Table 3: Averages and standard deviations of parameter errors of each diffusion model with corresponding synthetic datasets ([models] DM:Direct Infl. Model, IM:Indirect Infl. Model, [param. in DM/IM] p_i:external influence of individuals of type i, n_i:population of individuals of type i, [param. in DM] c_{ij}:internal influence of neighbors of type i on individuals of type j, [param. in IM] c_i:internal influence of a network of type i, b_{ij}:inter-network influence of a network of type i on a network of type j, [subscripts] n:News, s:SNS, b:Blog).

	News					SNS					Blogs				
DM	p_n	c_{nn}	c_{sn}	c_{bn}	n_n	p_s	c_{ns}	c_{ss}	c_{bs}	n_s	p_b	c_{nb}	c_{sb}	c_{bb}	n_b
AVG	8.7e-3	1.0e-2	2.3e-2	2.5e-2	9.0e+0	7.1e-3	4.1e-2	4.1e-2	3.1e-2	6.7e+0	4.3e-3	2.4e-2	2.6e-2	2.7e-2	9.4e+0
STD	7.5e-4	2.6e-2	2.2e-2	2.0e-2	4.8e+0	4.9e-4	2.5e-2	3.6e-2	2.4e-2	1.9e+0	5.2e-4	9.0e-3	1.7e-2	2.4e-2	4.4e+0
IM	p_n	c_n	b_{sn}	b_{bn}	n_n	p_s	c_s	b_{ns}	b_{bs}	n_s	p_b	c_b	b_{nb}	b_{sb}	n_b
AVG	1.7e-6	5.8e-5	2.4e-3	2.4e-3	9.5e-2	1.7e-5	1.4e-4	1.8e-3	2.1e-3	6.9e-2	1.6e-5	1.3e-4	2.3e-3	1.7e-3	2.8e-1
STD	9.5e-7	3.4e-5	1.4e-3	1.4e-3	4.8e-3	1.1e-6	1.9e-5	1.4e-3	1.4e-3	1.4e-2	2.0e-6	1.6e-5	1.4e-3	1.3e-3	3.7e-2

Table 4: Averages (standard deviations in parenthesis) of model fitting errors (RMSE) of each diffusion model with real datasets. (BM:Bass Model, DM:Direct Infl. Model, IM:Indirect Inf. Model)

BM	DM	IM
8.756e-2 (4.925e-2)	**6.233e-2** **(3.233e-2)**	8.274e-2 (4.558e-2)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}\sum_i \left(A_i(t)/n_i - P(a|i,t,\boldsymbol{\theta}_i)\right)^2}{3T}}, \qquad (19)$$

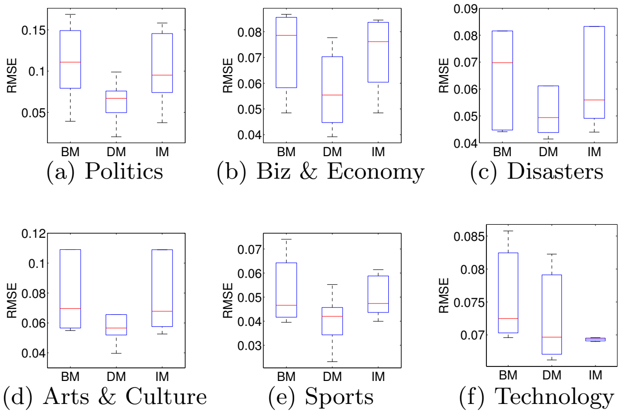
where n_i ($i = n, s, b$) denotes the population of each media type, and $P(a|i,t,\boldsymbol{\theta}_i)$ is the adoption probability $P(a|i,t)$ given the model parameters $\boldsymbol{\theta}_i$. Note that due to the parameter identification problem, where the same results are produced with different settings of parameters, we fix the power law coefficient α_i to be 2.5, whose value is typically reported to be in the range of $2 < \alpha < 3$ [9].

Table 2 shows the averages and standard deviations of model fitting errors (RMSE) of three diffusion models (BM, DM, and IM) for two groups of datasets. DM fitted best to the group of datasets generated by DM, while IM fitted best to the group of datasets generated by IM. Therefore, the proposed models are distinguishable and can be used to find hidden diffusion patterns. In both cases, the model fitting errors of DM and IM were always better than BM.

We also evaluated parameter errors as shown in Table 3. The RMSE and standard deviations of parameter errors are acceptable when compared to typical values ($p \approx 0.03$, $q \approx 0.3$ and $m \approx 3000$) [17] and show the feasibility of our models to recover parameters from datasets.

6.2 Experiments on Real Data

In Section 5, we described the data collection. From the largest connected document networks, we selected 35 news stories which have driven adoptions of at least 300 identified users across social media. Each news story is categorized by six topics by referring to the Wikipedia Current Events as shown in Table 1. The corresponding documents (650K) and users (94K) are extracted from the largest document network. With these 35 datasets, we examine how diffusion of news contents can be interpreted in terms of strength, direction, and direct/indirectness of influence between different types of social networks. Also, the variations of diffusion patterns across topics are investigated.



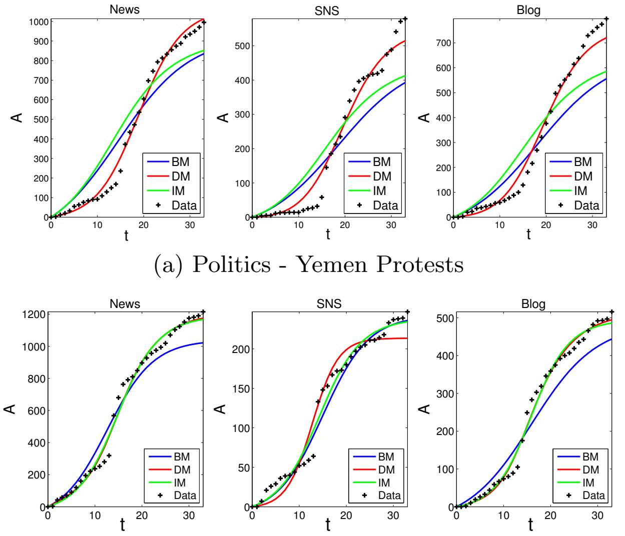
(a) Politics (b) Biz & Economy (c) Disasters

(d) Arts & Culture (e) Sports (f) Technology

Figure 3: Distribution of model fitting errors (RMSE) of BM, DM and IM with real datasets by six topics.

Direct/Indirectness of Influence: There is no ground truth of parameter values in the real datasets, so we fit the proposed models (DM and IM) and the baseline model (BM) using Nonlinear Least Squares (NLS) as in the experiments on synthetic datasets, and evaluate model fitting errors as shown in Table 4. Overall, due to noise in the real datasets, the averages of model fitting errors increased by one order of magnitude compared with those of the synthetic datasets in Table 2. However, our proposed models still perform better than BM, and in particular DM outperforms BM and IM by 29% and 25%, respectively, with more acceptable standard deviations. That is, diffusion of news contents in social media is better explained by direct than indirect influences between heterogeneous social networks. Also, analyzing diffusion within single social networks is not enough to fully describe its own diffusion because it neglects the effects of interactions with other social networks on diffusion. Thus, this result can be interpreted that real-world news diffusion in social media is attributed to globally and directly connected online social networks.

Diffusion Patterns by Topics: We classified the model fitting errors by topics as shown in Fig. 3 in order to see the variations of diffusion patterns across topics. Overall, most of the topics are better explained by DM, but the Tech-



(a) Politics - Yemen Protests

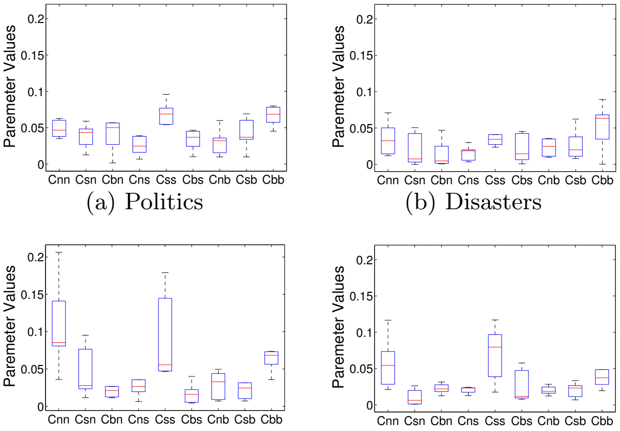
(b) Technology - Apple iPad for Education

Figure 4: Model fitting results with real datasets; representative cases of Politics and Technology topics are shown (BM: Bass Model, DM:Direct Influence Model, IM:Indirect Influence Model, A: cumulative adopters up to time t).

nology topic is evidently well traced by IM. This can be interpreted that online users are much more responsive to popular trends (a sudden increase of diffusion rate) of technology (e.g., Google technology trends) or product-related news (e.g., new iPad release) than other topics because IM traces well abrupt changes of diffusion rates in other networks.

In this regard we examine different diffusion patterns between the topics of Politics and Technology in more detail. Fig. 4 shows two representative cases of model fitting results from each topic. In the example of the Yemen protests in Fig. 4(a), the diffusion began to grow in news media first and then followed by blog and SNS media in this order. The growth rate was not rapid at the beginning, but the diffusion across all social media start to grow almost at the same time after two weeks. Without direct interactions across social media, such simultaneous growth unlikely happens. As the figure shows, BM has difficulty to follow such concurrent growth patterns without information about hidden connections between multiple social networks. IM better follows the growth than BM using the information of diffusion rate of other social media, but it is still not enough because the effect of coefficient b in IM is too small to trace the big jump.

On the other hand, the news about iPad as an educational device rapidly spreads from the beginning compared to the previous case, and thus there is enough time for IM to catch up the trend (still better than BM due to the coefficient b). DM also traces well, but diffusion trends across all social media are not synchronous as the case of the Yemen protests, which leads to the saturation of diffusion rate in SNS media by the larger coefficient c_{ss} than the other c_{ns} and c_{bs}. Overall, DM traces well synchronous diffusion across different types of social networks, while IM better catches the effect of rapid changes in diffusion rates of other networks in an asynchronous way.



(a) Politics

(b) Disasters

(c) Arts & Culture

(d) Sports

Figure 5: Distribution of the estimated parameter values with real datasets by topics (c_{ij} indicates the coefficient of DM which represents internal influence of neighbors of type i on the individuals of type j).

Heterogeneity versus Influence: Since there exist no ground truth of parameters for the real datasets, we analyze variations of inferred parameter values by different topics, instead of evaluating parameter errors. In Section 5.2, we examined the time-evolving heterogeneity of diverse network participation in news diffusion. The Politics and Disasters topics showed the highest relative entropy, while the Arts and Sports topics drove the lowest relative entropy. Accordingly, we compared the two pairs with distributions of the estimated parameter values of c_{ij} which indicates influence of media type i on other media type j, as shown in Fig. 5. As the figure shows, in general interactions within homogeneous social networks are stronger than ones between heterogeneous social networks, but they exhibit different levels of interactions by categories.

The topics of Politics (Fig. 5(a)) and Disasters (Fig. 5(b)) exhibit relatively even parameter value distributions, while the topics of Arts and Culture (Fig. 5(c)) and Sports (Fig. 5(d)) have relatively high medians in the coefficients for influence of same media types (c_{nn}, c_{ss} and c_{bb}) than others. Such different interaction patterns reflect the different levels of heterogeneity of populations in diffusion. That is, balanced interactions across social media in the cases of the Politics and Disasters topics likely drive evenly distributed populations of different types, i.e. high relative entropy. On the other hand, unbalanced interactions in the Arts and Sports categories likely come up with disproportionate population distribution, i.e. the lower relative entropy. Therefore, there exists dynamics of influence between heterogeneous social networks in accordance with topics of information.

Comparison of Outcomes with Related Work: As discussed earlier in Section 2, the authors of [18] discovered that political topics in Twitter are considerably driven by the external influence, while other topics such as entertainments are driven by internal communications. This outcome is consistent with our experimental results. That is, political topics, likely driving balanced interactions across different types networks, can be interpreted as high external influence from the aspect of single social platforms. The Arts and

Sports topics showed strong interaction patterns within homogeneous networks but unbalanced and weak interactions between heterogeneous networks, which can reflect internal communications within a single social platform driven by entertainment topics like in the case of [18]. In other words, topics of information affect spreading behaviors across different types of social networks, which enhances understanding diffusion phenomena within homogeneous networks.

7. CONCLUSION

By identifying real-world news stories from the Wikipedia Current Events and collecting the relevant hyperlink cascades across social media, we could observe the global diffusion patterns rather than local and site-specific diffusion. We proposed new conceptual frameworks of interpreting heterogeneous social networks in terms of direct and indirect influence. Such interpretation can provide a basis of underlying diffusion mechanisms in a wide range of diffusion space at a macro level. For realization of the concept, we accordingly proposed the Direct and Indirect Influence Models by considering two essential features, structural connectivity and heterogeneity of populations, at the same time. As a result, we generalized the Bass model into the dynamics of meta-populations with a probabilistic approach.

Experiments on both synthetic and real datasets showed the feasibility of the proposed models. Real-world news diffusion is found to be better explained by the direct influence model than the indirect one. There were also variations of diffusion patterns by topics. For instance, political and technological topics drove direct and indirect influences among different types of social media, respectively. That is, online users tend to be more responsive to rapid changes of technical trends. Such responses are well traced by the indirect diffusion model. On the other hand, political topics such as the Yemen revolution tend to spread concurrently across social media. Such phenomena unlikely happen without direct interactions between multiple social networks, which are well explained by the direct influence model.

Based upon the supportive evidences such as increasing relative entropy of populations in diffusion and their variations by topics, we examined the distributions of estimated parameter values with real datasets. Overall, interactions between the same media types are stronger than ones between different types. However, regarding the topics of Politics and Disasters, different social media tend to interact in a balanced way, while entertainment topics such as Arts and Sports exhibit stronger internal connections within the same types of social media like internal buzz, but showing unbalanced and weak interactions with other social networks. Such emergent phenomena can be explained with relative entropy. Balanced interactions naturally drive high relative entropy, and unbalanced interactions likely come up with relatively lower entropy of the participation of diverse social media.

We expect that the proposed models apply to a wider class of diffusion phenomena in diverse areas and provide a way of interpreting dynamics of meta-populations in terms of strength, direction, and direct/indirectness of influence. As future work, one possible topic is to improve the proposed models by using other properties of real-world social networks and to discover the evolving patterns between different kinds news contents at a macro level.

8. REFERENCES

[1] ICWSM'11 Dataset. http://www.icwsm.org/data/.

[2] Wikipedia Current Events in January, 2011. http://en.wikipedia.org/wiki/January_2011.

[3] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *WI*, pages 207–214, 2005.

[4] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[5] F. Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[6] F. Bass. Comments on "a new product growth for model consumer durables": The Bass model. *Management Science*, pages 1833–1840, 2004.

[7] F. Bass, T. Krishnan, and D. Jain. Why the Bass model fits without decision variables. *Marketing Science*, pages 203–223, 1994.

[8] M. Cha, J. Pérez, and H. Haddadi. Flash floods and ripples: The spread of media content through the blogosphere. In *ICWSM*, 2009.

[9] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501. ACM, 2004.

[11] A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

[12] M. Kim, L. Xie, and P. Christen. Event diffusion patterns in social media. In *ICWSM*. AAAI, 2012.

[13] V. Kumar and T. Krishnan. Multinational diffusion models: An alternative framework. *Marketing Science*, pages 318–330, 2002.

[14] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506. ACM, 2009.

[15] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM*, 2007.

[16] M. Luu, E. Lim, T. Hoang, and F. Chua. Modeling diffusion in social networks using network properties. In *ICWSM*, 2012.

[17] V. Mahajan, E. Muller, and F. Bass. Diffusion of new products: Empirical generalizations and managerial uses. *Marketing Science*, 14(3):G79–G88, 1995.

[18] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD*, pages 33–41. ACM, 2012.

[19] M. Newman. *Networks: an introduction*. Oxford University Press, 2010.

[20] W. Putsis Jr, S. Balasubramanian, E. Kaplan, and S. Sen. Mixing behavior in cross-country diffusion. *Marketing Science*, pages 354–369, 1997.

[21] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics. In *WWW*, pages 695–704. ACM, 2011.

[22] F. Stutzman, J. Vitak, N. B. Ellison, R. Gray, and C. Lampe. Privacy in interaction: Exploring disclosure and social capital in facebook. In *ICWSM*, 2012.

[23] J. H. Zar. *Biostatistical Analysis*. Prentice Hall, 2010.