Unexpected Relevance: An Empirical Study of Serendipity in Retweets

Tao Sun*
School of EECS
Peking University
suntao@net.pku.edu.cn

Ming Zhang
School of EECS
Peking University
mzhang@net.pku.edu.cn

Qiaozhu Mei
School of Information
University of Michigan
qmei@umich.edu

Abstract

Serendipity is a beneficial discovery that happens in an unexpected way. It has been found spectacularly valuable in various contexts, including scientific discoveries, acquisition of business, and recommender systems. Although never formally proved with large-scale behavioral analysis, it is believed by scientists and practitioners that serendipity is an important factor of positive user experience and increased user engagement. In this paper, we take the initiative to study the ubiquitous occurrence of serendipitious information diffusion and its effect in the context of microblogging communities. We refer to serendipity as *unexpected relevance*, then propose a principled statistical method to test the unexpectedness and the relevance of information received by a microblogging user, which identifies a serendipitous diffusion of information to the user. Our findings based on large-scale behavioral analysis reveal that there is a surprisingly strong presence of serendipitous information diffusion in retweeting, which accounts for more than 25% of retweets in both Twitter and Weibo. Upon the identification of serendipity, we are able to conduct observational analysis that reveals the benefit of serendipity to microblogging users. Results show that both the discovery and the provision of serendipity increase the level of user activities and social interactions, while the provision of serendipitous information also increases the influence of Twitter users.

Introduction

"Serendipity is looking in a haystack for a needle and discovering a farmer's daughter."

Julius Conroe, Jr.

Conventionally, serendipity is described as "pleasant surprise" (Golin 1957), "unintended finding" (Andel 1994), or "accidental discovery" (Roberts 1989). Numerous examples in the literature show that serendipity plays an important role in the innovations of arts, science, and technology, such as the discovery of Teflon, Velcro and sugar substitutes (Roberts 1989). In the context of modern information systems, serendipity has been demonstrated to be useful as well. For example, serendipitous recommendations help users find surprisingly interesting items and eventually increase the volume of transactions in recommender systems (Kawamae 2010). The exposure to serendipity can also lead users to new, unanticipated outcomes in an information retrieval system (Foster and Ford 2003). To summarize, there is a common belief that serendipitous discovery is strongly associated to positive user experience and increased user engagement in information systems, although such effect has never been quantified with scale.

Even though serendipity has not been formally studied with large-scale behavioral analysis, practitioners have moved forward towards various system designs that enhance serendipity in various contexts. Examples include Google's attempt in its theoretical serendipity engine[1] and eBay's test in serendipitious shopping (Woyke 2011). A formal study of serendipity with large-scale behavioral analysis will not only provide a solid rationale of these explorations of enhancing serendipity, but also shed light on how it can be achieved.

The popularity of microblogs has created an unprecedented opportunity to observe and analyze user behaviors at a very large scale. A typical microblogging user can follow anyone without acquiring the consensus of the followee, forming many weak ties which in turn diffuses novel information (Granovetter 1973). Meanwhile, one user's updates will be automatically pushed to his followers, which raises the followers' chances to encounter unexpected and interesting content. Reading tweets in a timeline is more comparable to reading headlines in a newspaper, which has been described as an aggregation of serendipitous information (Toms 2000). Under both circumstances, readers can easily access unanticipated information they are not in quest of. Thus, microblog is a natural habitat of serendipitous discoveries, due to the asymmetric structure of relationship, the high volume of information flow, the mechanism of hashtags (Kop 2012; Conover et al. 2011), and the indirect media exposure (An et al. 2011).

We anticipate that microblogging communities provide a suitable context to observe not only the presence of serendipity, but also the effect of serendipity. Consider a pair of Twitter users Annie and Bill. Bill has always been tweeting technology news, and Annie follows him mainly

*The majority of this work was done when Tao Sun was visiting the University of Michigan sponsored by China Council Scholarship.

[1] http://techcrunch.com/2010/09/28/eric-schmidt-future-of-search/

for the latest IT trends. One day Bill happened to tweet about a new medication that helps to resolve the symptoms of motion sickness. Coincidentally, Annie had been suffering from travel sickness and was in desperate need of a remedy. Thus, the tweet turns out to be a serendipitous discovery by Annie: it is unexpected from Bill (he has never posted about motion sickness before) and relevant to Annie (she is in need of a cure). With the example, one would anticipate that such an unexpected discovery would enhance Annie's experience of the community, and increase her engagement to the context. Besides, this surprising satisfaction of Annie's information need is likely to cause her to interact more frequently with Bill, and possibly to reply and retweet more actively.

While individual examples are easy to identify, a large-scale quantitative analysis of serendipity is challenging. The identification of serendipity requires a formal test of unexpectedness (e.g., the medication tweeted by Bill), which refers to unpredictability and uncertainty (André, Teevan, and Dumais 2009). It also requires a way to assess the relevance/usefulness of a tweet to the receiver (e.g., Annie). These tasks are non-trivial given the sparseness of information to estimate Bill and Annie's preference and information needs.

In this paper, we propose a formal definition that characterizes serendipity in the context of information diffusion. Although the definition is motivated from microblogging, it can be applied in a variety of contexts such as recommender systems and retrieval systems. Then we proposed a formal statistical method based on likelihood ratio test to examine the unexpectedness, the relevance, and serendipity in a diffusion process from a source to a receiver.

With the proposed method, we are able to quantitatively examine serendipity in Twitter and Weibo, which are the leading microblogging sites in English and Chinese respectively. Our experiments reveal a high ratio of serendipitous diffusion of information though retweeting. Serendipity presents in 27% of retweets in Twitter, and in 30% of retweets in Weibo. Based on the statistical study, we further investigate whether serendipity can lead to enhanced user activities and engagement. Specifically, we compute correlations between serendipity and the change of user status and behaviors. Our results show that serendipity does play a positive role in increasing the level of user activities and social interactions. More specifically, the more serendipitous content a user discovers, the more active and sociable he will become; the more serendipitous content a user provides, the more active and influential he will become.

To the best of our knowledge, this is the first quantitative study of the phenomenon of serendipity with large-scale behavioral analysis. The major contributions are summarized as below.

(1) We formulate and characterize the phenomenon of serendipity that fits into different types of scenarios.

(2) We propose a principled method based on statistical hypothesis tests to identify serendipitous diffusion of information in microblogging communities.

(3) We provide quantitative estimates that reveal the presence of serendipity in retweeting behaviors, and demonstrate the positive effect of serendipity on user engagement and social status.

Related Work

The significance of serendipity has long been recognized, e.g., in scientific knowledge such as the discovery of penicillin, safety glass, Newton's theory of gravitation, and the discovery of DNA (Roberts 1989). Intuitively, serendipity leads to the "Aha!" experience (McCay-Peet and Toms 2011),when a previously incomprehensible problem or concept becomes suddenly clear and obvious.

Recently, in the context of modern information systems, it has been demonstrated that serendipity improves user satisfaction (Zhang et al. 2012), delights users by helping them discover new items (Bellotti et al. 2008), reinforces an existing direction or leads to a new direction in information seeking (Foster and Ford 2003). Serendipity has also found its way into the evaluation metrics of recommender systems. Ziegler et al. (2005) argued that beyond accuracy, there are other important traits of user satisfaction such as serendipity, which is inversely proportional to popularity. Abbassi et al. (2009) viewed serendipitous recommendations as Outside-The-Box($\mathcal{OTB}$) items, which is related to unfamiliarity. Zhang et al. (2012) used serendipity as a performance metric, where *unserendipity* was assessed by calculating similarity between one user's history and a recommendation list. Based on the assumption that unexpectedness is low for items predicted by a primitive prediction method, serendipity was also measured by comparing to items generated by prediction models (Murakami, Mori, and Orihara 2008; Ge, Delgado-Battenfeld, and Jannach 2010).

All the metrics of serendipity discussed in recommender systems are employed to assess the system usability and user experience at an abstract level, rather than to identify whether a specific item is serendipitous or not. Although some could be modified to identify serendipitous items, they all heavily rely on either heuristic thresholds such as $\mathcal{OTB}$ness scores (Abbassi et al. 2009), or additional prerequisite such as a primitive prediction model (Murakami, Mori, and Orihara 2008). A principled model of serendipity and a systematic way to identifying serendipitous information are needed.

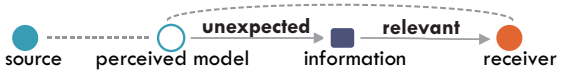
In the context of microblogging, serendipity has also been studied by small-scaled, survey-based behavioral analysis. Most of the measurements make use of surveys or interviews that invite users to manually check serendipity and ask them to what extent serendipity is helpful (Herring et al. 2011; Piao and Whittle 2011). For example, Piao and Whittle (2011) identified serendipitous terms by manually rating tweets with two metrics interesting and surprising. However, such methods fail to provide an automatic method to identify serendipitous content, and certainly do not scale to the real context of microblogging (e.g., 200 million active users and 400 million tweets per day[2]).

The concept of serendipity is also related to, but substantially different from concepts like novelty (Granovetter 1973) and diversity (Reagans and Zuckerman 2001), which

[2]http://techcrunch.com/2012/09/19/tweets-on-twitter/

(a) Information diffusion at the microscopic level

(b) Information diffusion incorporated with perceived model



(c) A serendipitous diffusion of information

Figure 1: The illustration of information diffusion, a perceived model, and serendipity

are well studied in literature. The difference will be elaborated in the next section.

Definition

The definition of serendipity varies in literature, such as *pleasant surprise* (Golin 1957), *unintended finding* (Andel 1994), or *accidental discovery* (Roberts 1989). What's in common among all these narrative definitions of serendipity is the *surprising* nature of serendipitous events. In other words, a serendipitous discovery has to be *unexpected*. Such an aspect is adopted in most work that applied serendipity to recommender systems. For example, Zhang et al. (2012) referred to serendipity as the *unusualness* or *surprise* of recommendations.

Another implicit yet crucial aspect in these definitions is the notion of *usefulness*. An unexpected discovery is classified as serendipitous only if it is considered to be pleasant, interesting, or useful. In a recent workshop of understanding the nature of serendipity (Makri and Blandford 2011), it is the consensus that besides the *unexpected nature* there must also be a "*clear, positive outcome*" in the classification of serendipitous events. Based on these intuitions, below we formally define serendipity in the context of information diffusion.

Serendipitious Diffusion

From a microscopic viewpoint, the diffusion of a piece of information was determined by three dimensions: the *source* of the information, the target or *receiver* of the information, and the piece of *information* being delivered, as illustrated in Fig.1.a.

Repeatedly receiving information from one source, the receiver is likely to develop a perception about the source. Based on this *perceived model* of the source, the receiver forms an expectation about the next piece of information to be sent from the source, which can be concerned with the frequency of diffusion, the type of diffusion, or the topic of information. Certainly, such an expectation may be biased by the limited observation about the source. In other words, the perceived model serves as an approximation of the invisible ground truth about the source, and captures the receiver's expectation about the source.

Then, the process of information diffusion from the receiver's point of view can be re-illustrated as in Fig.1.b. Different from the model in Fig.1.a., all the receiver's understanding about the source is now through the perceived model. That is to say, from the receiver's perspective, the information diffused to him is generated from the perceived model about the source, rather than directly from the source. With such a model of diffusion, we then formally define the unexpectedness and the usefulness of the information being diffused, which lead to the definition of serendipity.

Unexpectedness, Relevance & Serendipity

Let s be a source of information, r be a receiver, and x be a piece of information diffused from s to r. Based on the model in Fig.1.b., we define φ_s as the perceived model derived by r. If the receiver perceives that x is likely to be generated from φ_s, then the piece of information being diffused is *expected* by r; otherwise it is *unexpected* by r. Formally, we have

Definition 1 UNEXPECTEDNESS. If a piece of information x is sufficiently divergent from the perceived model φ_s, x is defined as unexpected from the source s by the receiver r.

Both the piece of information and the perceived model can be represented with a rich set of attributes, e.g., time of infection, type of diffusion, and topic of information. In this study we only focus on the unexpectedness with regard to the topical content, but the definition is general enough to capture other aspects of unexpectedness. In practice, the method to assess the divergence of x and φ_s can vary. If φ_s is represented as a probabilistic model, then the unexpectedness of x can be assessed by how *unlikely* x is generated from φ_s, or how unlikely x can be predicted by φ_s.

Comparing to the unexpectedness, the usefulness or positiveness of information is even more difficult to define. Although naturally every receiver has his own model to assess usefulness, in reality such a model is hard to observe. In some context, we may view every successful diffusion as useful, such as the adoption of new technique or product. But it is subtle in microblogs, e.g., whether a tweet is received by one user is hard to identify (i.e., he might have never read a tweet even if it appeared in his timeline). Even retweeting does not always indicate the content being retweeted is useful or interesting. It could simply be a behavior to compliment a friend, a strategy to maximize influence, or an approach to label tweets for future personal access (Boyd, Golder, and Lotan 2010). Despite this difficulty, we make a practical assumption and assess the usefulness or interestingness of a piece of information by how *relevant* it is to the user's interest. The interest of a user can be estimated from what he has posted. Therefore, a tweet is defined as *useful* to or *positively received* by a user if he retweets it and the tweet is *relevant* to what he usually tweets about. This assumption does limit the scope of our study to retweeting behaviors. Formally, we have

Definition 2 RELEVANCE. *Let φ_r be a model of user r's preference of information. Given a piece of information x received by r from any source, if x is sufficiently close to φ_r, we define x as relevant to the user r.*

In practice, we first need to estimate the model φ_r. Given observations of what receiver r has posted, φ_r can be estimated from all the tweets written by r. This is similar to the model of user preference in recommender systems. The definition also requires a decision whether a piece of information is indeed "received". As stated above, we use retweeting as one explicit signal that a user "receives" a tweet.

The above definitions of *unexpectedness* and *relevance* are general enough to be applied to various scenarios. For example, in a recommender system, the source is the system itself, the receiver is the user and the information is a recommended item. The item is unexpected if it is divergent from what the system commonly recommends to a user (e.g., a liberal blog recommended to republicans), and relevant to the user if it is close to what the user has clicked/purchased in the past. While in a retrieval system, similarly, the source is the system itself, the receiver is the user, and the information is a returned result. The result is relevant to a user if it is close to the query or to what he has clicked/browsed. And it appears to be unexpected if it is sufficiently divergent from what the system usually returns (e.g., a paper written by computer scientists returned by PubMed, the search engine of biomedical literature).

In the particular scenario of microblogging, one thing worth particular attention is the availability of multiple *contexts* that could explain the piece of information being diffused to a user other than his perceived model φ_s. For example, if Justin Bieber posts "Happy New Year" (the information) on Jan 1, it is unexpected from Justin Bieber (source) if he never sent new year greetings, but it is not unexpected from a different context — the time of the year. Other than the context of time, one can imagine other contexts such as the geographical location, particular events and trends. A tweet may be unexpected from one context but expected from other contexts. Similar to φ_s, we introduce a model for every possible context, such as φ_t for the time context.

We assume that there are a finite number of contexts in microblogging, and a tweet is generated from a mixture of the contexts. Formally, a tweet x is generated from a mixture model $\Phi = \sum_i \lambda_i \varphi_i$, where the corresponding source φ_s is one of the mixture components ($\varphi_s \in \{\varphi_i\}$) and λ_i is the weight coefficient of the i^{th} context. We have $0 \leq \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$. When φ_s is the only component of the mixture model that has a positive mixture weight (i.e., $\Phi = \varphi_s$), we say the tweet x is generated from φ_s, or the tweet is expected from s. When Φ is sufficiently close to φ_r, we say the tweet x is relevant to r.

The definitions of unexpectedness and relevance jointly define serendipity in information diffusion. Intuitively, serendipity is a subjective phenomenon implying an *unexpected* discovery of information that is *relevant* to the user's need or preference, which is along the line with other definitions in literature (Foster and Ford 2003; De Bruijn and Spence 2008; Jarvis 2010; Piao and Whittle 2011; Case 2012; Kop 2012). As illustrated in Fig.1.c, we have

Definition 3 SERENDIPITY. *Given a piece of information x being diffused from a source s to a receiver r, and a perceived model φ_s of s developed by r, the diffusion of x is serendipitous to r if and only if x is unexpected from s (i.e., φ_s) and x is relevant to r.*

Note that the concept of serendipity is different from other concepts like *novelty* and *diversity*. E.g., novel information is more likely to be diffused through weak ties than strong ties (Granovetter 1973). However, the novelty of information does not imply surprisingness, which is a key prerequisite of serendipity. For example, a tweet from Justin Bieber about his new album is novel, but not surprising. A real surprise is when Justin Bieber tweets about data mining algorithms. Serendipity is also related to the notion of diversity (Reagans and Zuckerman 2001). However, the diversity of information does not imply the relevance or usefulness, which is the other key criterion of serendipity. For example, receiving a tweet about "data mining" provides diversity to an artist, but may be considered as unanticipated disturbance. The definition of serendipity integrates the surprisingness (unexpectedness) and usefulness (relevance), which provides a novel perspective of understanding the characteristics of information diffusion.

With the general definition of serendipity, we propose statistical methods to formally identify serendipitous diffusion of information in microblogging.

Identification Method

Traditional studies of serendipity identification usually employ survey-based approaches, where human subjects are asked to label serendipitous events explicitly (Herring et al. 2011; Piao and Whittle 2011; Bellotti et al. 2008). These approaches typically do not scale up. In order to quantify serendipity at a large scale, we need a principled method to automatically identify serendipitous events. Intuitively, once the diffused information x, the perceived model of the source φ_s, and the preference of the receiver φ_r are mathematically formulated and estimated, we simply need a way to assess the divergence between x and φ_s and the closeness between x and φ_r.

There are various ways to represent x, φ_s and φ_r, such as vector space models or statistical language models. One possibility is to represent all the models as term vectors, and apply vector-based similarity/distance measures (e.g., the cosine similarity) to assess the divergence and closeness. We do not pursue such methods because of two reasons. Firstly, tweets are too short and the dimensionality of vocabulary is too high. The effectiveness of distance measures like cosine similarity is largely compromised due to data sparsity. Secondly, even if advanced IR methods such as text normalization (Han, Cook, and Baldwin 2013) can be employed to combat sparsity, decision-making methods using distance measures still rely on arbitrary thresholds. One has to find magic thresholds of similarity that distinguish unexpectedness from expectedness, and relevance from irrelevance. There is no guidance on how to select such thresholds to identify serendipity.

In our study, we propose a principled method that utilizes likelihood ratio test to identify unexpectedness and relevance. Such a statistical method does not rely on arbitrarily selected thresholds.

The Likelihood Ratio Test

Likelihood ratio test is a statistical hypothesis test for model selection by comparing the goodness of fit (Neyman and Pearson 1933). Its efficiency has been proved in numerous scientific experiments, such as DNA substitution (Posada and Crandall 1998) and phylogenetic analysis (Huelsenbeck, Hillis, and Nielsen 1996). Let Θ be the entire parameter space, and the model being tested be an instantiation from a parameterized family of statistical models with parameter θ ($\theta \in \Theta$). The null hypothesis H_0 restricts the parameter θ to Θ_0 that is a subset of Θ, and the alternative hypothesis H_1 restricts the parameter θ to Θ_1 that is another subset of Θ. Let $L(\theta|x)$ be the likelihood that the observed data x is generated from the model with parameter θ. The test statistic is $D = -2ln(\Lambda)$, where Λ is the likelihood ratio being

$$\Lambda = \frac{max[L(\theta|x), \theta \in \Theta_0]}{max[L(\theta|x), \theta \in \Theta_1]} \tag{1}$$

In Equation 1, Λ is the comparison between the maximum likelihood under H_0 with that under H_1. A small Λ indicates that the alternative model can explain data better than the null model. When the competing models are *nested*, or the null model is a special case of the alternative model (i.e., $\Theta_0 \subseteq \Theta_1$), D converges in distribution to a χ^2-distribution with degree of freedom (*df*) that equals to the difference in the dimensionality of Θ_0 and Θ_1 (Wilks 1938).

The Test of Unexpectedness

With the general description of the likelihood ratio test, now we describe our method to test whether one piece of information is unexpected from the source.

Here, our goal is to examine whether the diffused information x is expected from φ_s. Receiver r views x as expected if x is likely to be generated from φ_s, and unexpected if x is substantially more likely to be generated from a model different from φ_s.

As discussed in the Definition section, one can assume that x is generated from a mixture model of multiple contexts, $\Phi = \sum_i \lambda_i \varphi_i$, with φ_s as one of the component models. Therefore, the test is essentially to find out whether the tweet is more likely to be explained by the perceived model of source alone, or more likely to be explained by the mixture of other contexts such as the time and the general background. In this study, we instantiate the models of all the contexts (φ_i) as unigram language models (i.e., multinomial distributions of bag of words), which is a common practice in information retrieval (Ponte and Croft 1998; Manning, Raghavan, and Schutze 2008).

Formally, to verify the unexpectedness of x from φ_s, we have H_0 being "x is generated from φ_s along", whereas H_1 being "t is generated from an alternative model". The alternative model is a mixture model of multiple contexts Φ, which includes φ_s as a mixture component. This ensures that the null model and alternative models are nested. If H_0 is rejected at a predefined level of confidence, we can conclude that x is unexpected from model φ_s. More specifically, it means the true model that generated x has at least one different component model besides φ_s. In other words, at least part of x is unexpected from the source s.

We use θ to denote the model parameters of Φ, i.e., $\theta = \{\lambda_i\}_{i=1...k}$. The null hypothesis and the alternative hypothesis are then given as below.

H_0: $\theta \in \Theta_0 = \{\theta|\lambda_s = 1; \forall i \neq s, \lambda_i = 0\}$.
H_1: $\theta \in \Theta_1 = \{\theta|\forall i, 0 \leq \lambda_i \leq 1; \sum_i \lambda_i = 1\}$.

To conduct the likelihood ratio test, we need to compute $max[L(\theta|x), \theta \in \Theta_1]$. In fact, all alternative models with the parameter space Θ_1 are certain mixture of contextual models, $\Phi = \sum_i \lambda_i \varphi_i$, with various weighting coefficients λ_i. The value of $max[L(\theta|x), \theta \in \Theta_1]$ is computed by finding the optimal coefficients λ_i's that maximize the log-likelihood $\log p(x|\Phi)$. We make the bag-of-words simplification that

$$p(x|\Phi) = \Pi_{w \in x} p(w|\Phi)^{c(w;x)} = \Pi_{w \in x} (\sum_i \lambda_i p(w|\varphi_i))^{c(w;x)} \tag{2}$$

where $p(w|\varphi_i)$ is the probability of the word w being generated by the language model φ_i of the i^{th} context, and $c(w;x)$ is the count of word w in x. $p(w|\varphi_i)$ can be estimated using a maximum likelihood estimator to fit the data observed in the i^{th} context, smoothed by the Jelinek-Mercer smoothing method (Zhai and Lafferty 2001). Specifically, with β being a global smoothing parameter satisfying $0 \leq \beta \leq 1$, we have:

$$p(w|\varphi_i) = (1 - \beta)p(w|D_i) + \beta p(w|C). \tag{3}$$

where D_i is the entire set of information observed in i^{th} context and C is a large collection of tweets as the background. With $c(w; D_i)$ denoting the count of word w in D_i, $p(w|D_i)$ and $p(w|C)$ are simply maximum likelihood estimates from words in D_i and C, e.g.,

$$p(w|D_i) = \frac{c(w; D_i)}{\sum_w c(w; D_i)}, \tag{4}$$

A local optimal of $p(x|\Phi)$ can be found by an EM algorithm that applies the following iterative updating process until converging:

$$\text{E-step:} \quad z_{w,j} = \frac{\lambda_j p(w|\varphi_j)}{\Sigma_i \lambda_i p(w|\varphi_i)} \tag{5}$$

$$\text{M-step:} \quad \lambda_j = \frac{\Sigma_w z_{w,j} c(w;x)}{\sum_w c(w;x)} \tag{6}$$

The Test of Relevance

The next task is to identify whether the diffused information x is relevant to the receiver r. We first represent the preference model of r, φ_r, as a language model estimated from all the information r has posted. The model is smoothed with the same process above.

We then present a different test, with the null hypothesis being "x is irrelevant to φ_r", and the alternative hypothesis

being "x is relevant to φ_r". Note it is a stronger statement that an observed piece of information is "expected" from a model than that an observation is "relevant" to a model. E.g., a tweet "how data mining helps election" is *relevant* to all users who tweeted about election, whereas not *expected* from those who tweeted "election" but never "data mining". Therefore, we define x as irrelevant to φ_r if x is not even partially generated from φ_r, and relevant otherwise. More specifically, suppose x is generated from Φ being a mixture model with multiple contextual components. If φ_r is not one of them (or $\lambda_r = 0$), we define x as irrelevant to φ_r. Formally, we conduct a likelihood ratio test as below, where H_0 becomes "x is generated from a mixture model *excluding* φ_r", and H_1 becomes "x is generated from a mixture model *including* φ_r."

$$H_0\text{: } \theta \in \Theta_0 = \{\theta | \forall i, 0 \leq \lambda_i \leq 1; \textstyle\sum_i \lambda_i = 1; \lambda_r = 0\}$$
$$H_1\text{: } \theta \in \Theta_1 = \{\theta | \forall i, 0 \leq \lambda_i \leq 1; \textstyle\sum_i \lambda_i = 1; \lambda_r \neq 0\}$$

Similar to the test of unexpectedness, the test statistic $D = -2ln(\Lambda)$ is computed by finding the optimal λ's that maximize $L(\theta|x)$ under the corresponding constraints, with the EM algorithm in Equ 5 and Equ 6. H_0 is rejected when D exceeds the critical value at predetermined alpha level of significance.

The Identification of Serendipity

Based on Definition 3, for a source s, a receiver r and a piece of information x, if x passes both the test of unexpectedness and the test of relevance, we have a certain level of confidence to conclude that x is a serendipitous piece of information diffused from s to r.

It is particularly noteworthy that, in the definition of Φ in this section, we do not specify the selection of component models φ_i other than φ_s or φ_r. The specific contexts will be selected when applied to real scenarios. Given the principled hypothesis tests, below we present a large-scale quantitative analysis of serendipity in microblogs such as Twitter and Weibo.

Experiments

Data Collection

With the statistical tests presented above, we examined serendipity with large-scale datasets collected from two different microblogging sites — Twitter and Weibo, which represent the most popular microblog communities in English and Chinese, respectively.

In Twitter, we collected data through Twitter Streaming API with the "Gardenhose" access level from 7/9/2011 to 11/30/2011. Gardenhose returns a random sample of all public statuses, accounting for approximately 10% of public tweets. Next, we randomly sampled $100,000$ users who had tweeted during the period, and extracted all the *retweets* during this period that involved the sampled users either as the sources (who are retweeted) or the receivers (who retweet), denoted as the TWITTER RT SAMPLE. As stated above, retweeting is an explicit signal that a tweet is actually "received" by a user, thus denoting a successful diffusion of information. We also extracted all the *tweets* written by all the sources and receivers involved in the TWITTER RT SAMPLE, all together $181,743$ valid users. These tweets were used to compute the preference models of the sources and the receivers.

Table 1: Statistics of the *sampled* users, excluding their ego-networks and public timeline

	Users	**RTs**	**Tweets**
Twitter RT Sample	100,000	587,021	2,723,595
Weibo RT Sample	2,000	67,564	206,528

In Weibo, due to the lack of streaming API, we extracted data through a Weibo API that is akin to Twitter Search API. Firstly, we randomly sampled a set of $2,000$ users who posted in December 2011, and crawled their ego networks (friends and followers). Then, we collected tweets posted by all the users above from 1/12/2012 to 6/30/2012, including the sampled users and users in their ego networks. It helped us extract retweets and identify who retweeted whom, because the Weibo API only gives the original author despite the long chain of retweets. We collected a total number of $514,170$ unique users from the ego networks of the $2,000$ seedling users. Similar to Twitter, all the retweets involving the $2,000$ sampled users were denoted as the WEIBO RT SAMPLE. During the same time period, we also crawled tweets from the public timeline, which displays the latest tweets from users sitewide. With an enterprise access level, about $2,000,000$ tweets were crawled from public timeline per day.

Table 1 lists the statistics of *the sampled users* from Twitter and Weibo, without counting in the statistics from the receivers or sources involved in the RTs, and the public timeline.

The Presence of Serendipity

Model Construction

To instantiate the mixture model Φ, we consider three contexts: the user himself, the time, and the background messages. For clarity, we refer to the three contextual models as the *user model*, the *temporal model* and the *background model*, respectively. The user model tells the user's preferences and interests. The temporal model captures the trending topics. The background model captures the public interest from a longer period of time, which also helps the estimation of the background word distributions with the purpose to combat sparsity (Lin, Snow, and Morgan 2011).

We assume the perceived model of one information source is shared by all users. We are however aware of heterogeneous perceptions, where different receivers have distinct perceived models. But it requires a comprehensive study on user behaviors such as login frequency, dwell time and reading habits, which is beyond the scope of this paper. And we leave it as one of our future directions. Considering the temporal dynamics (Yang and Leskovec 2011) and the evolving characteristic in perceived models, we update the perceived model of a source on a daily basis. On a specific day d, u's user model φ_u is estimated from all the tweets he posted before day d (excluding d). The temporal model φ_d

is estimated from all the tweets posted by *all* users on the previous day of d. The background model φ_b is estimated from all the tweets posted by all users before day d. Each model itself is smoothed by JM smoothing as Equation 3.

Hypotheses

In the experiment, we formally test the presence of serendipitous information diffusion in microblogging. We aim to investigate whether microblogging users can discover serendipity from another user, and whether microblogging users can provide serendipity to other users. In a diffusion of information in microblogging, a pair of users u and v behave as the information receiver and the information source, as an instantiation of Figure 1.c. In microblogging, as discussed above, even if u follows v, it is not guaranteed that u reads every single tweet posted by v. Therefore, we limit the study of information diffusion within *retweets*, which ensure that the tweets are successfully received by the receivers.

Let x be a retweet that u reposted from v. By Definition 3, a serendipitous retweet x should be *unexpected* from source v and *relevant* to receiver u. Let λ_u, λ_v, λ_t, λ_b be the weights of φ_u, φ_v, φ_t and φ_b in the mixture model Φ, which represent the context model of the receiver, the source, the temporal trends, and the background, respectively. Based on the statistical tests elaborated in the section of Identification Method, our goal is to test whether x is unexpected from φ_v and whether x is relevant to φ_u. The test of whether x is unexpected from φ_v is specified as:

$$H_0\text{: } \theta \in \Theta_0 = \{\theta | \lambda_v = 1; \forall i \neq v, \lambda_i = 0\}$$
$$H_1\text{: } \theta \in \Theta_1 = \{\theta | \lambda_v + \lambda_t + \lambda_b = 1; \forall i \neq v, t, b, \lambda_i = 0\}$$

The test of whether x is relevant to φ_u is specified as:

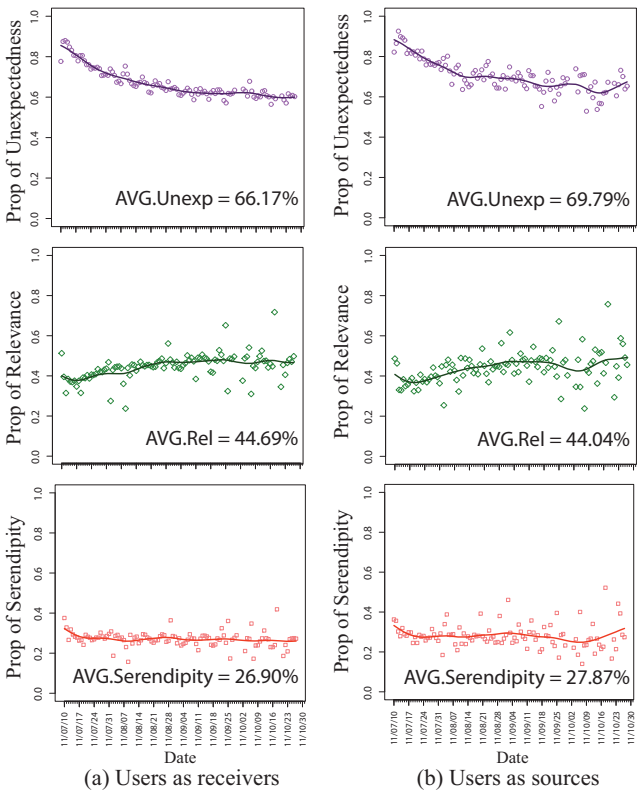
$$H_0\text{: } \theta \in \Theta_0 = \{\theta | \lambda_t + \lambda_b = 1; \forall i \neq t, b, \lambda_i = 0\}$$
$$H_1\text{: } \theta \in \Theta_1 = \{\theta | \lambda_u + \lambda_t + \lambda_b = 1; \forall i \neq u, t, b, \lambda_i = 0; \lambda_u \neq 0\}$$

For every retweet in TWITTER RT SAMPLE and WEIBO RT SAMPLE, we conducted the two Likelihood ratio tests. We excluded retweets whose receivers and sources have not posted before, because it is impossible to construct the user models. For each retweet, we constructed φ_u, φ_v, φ_t and φ_b, estimated λ's with the EM algorithm to calculate the maximum likelihood as Equ 5 and Equ 6, and then computed the value of test statistics as Equation 1. The df in the test of unexpectedness is 2, and df in the test of relevance is 1.

Results

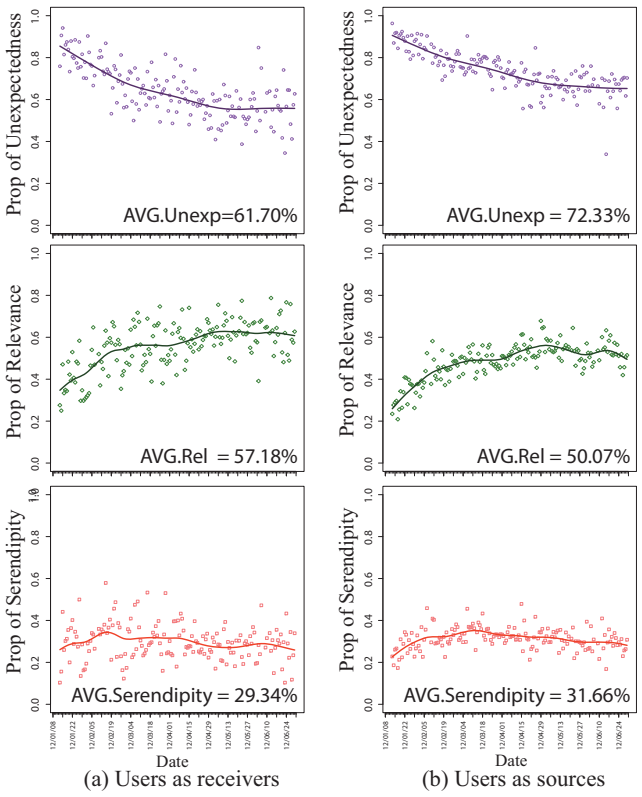
We are then able to conduct the tests of serendipity in Twitter during a time period from 7/10/2011 to 10/30/2011, and hold out the data from 11/1/2011 to 11/30/2011 to study the effect of serendipity which will be discussed shortly. In Weibo, we conducted hypothesis tests in a time span from 1/13/2012 to 6/30/2012. Retweets on the very first day in the collected data (7/9/2011 in Twitter and 1/12/2012 in Weibo) were excluded, because there is no previous observation to construct the user models. In all likelihood ratio tests, the level of significance α equals 0.05.

The results of the statistical tests from the TWITTER RT SAMPLE are shown in Fig. 2; and the results from the



(a) Users as receivers (b) Users as sources

Figure 2: Results in TWITTER RT SAMPLE



(a) Users as receivers (b) Users as sources

Figure 3: Results in WEIBO RT SAMPLE

WEIBO RT SAMPLE in Fig. 3. A microblog user can behave in a retweet either as the receiver or the source. Thus, we distinguish the two roles in our tests. In Fig. 2 and Fig. 3, the dates are displayed in x-axis, and the "Prop of Unexpected (Relevant, and Serendipity)" in y-axis is calculated as the number of unexpected (relevant, and serendipitous) retweets divided by the total number of retweets tested. This proportion is calculated independently day by day.

As we can see, in both Twitter and Weibo, the proportion of retweets identified as unexpected decreases fast at first, then converges to saturate. It is because the more tweets observed from the user, the less likely that his next tweet is unexpected. Similarly, the proportion of relevant tweets increases at first and gradually saturates at a certain level. It is because the more tweets observed from the user, the less likely that a relevant tweet will be misclassified as irrelevant. Observations from Weibo present similar trends, as shown in Fig. 3. By comparing Fig. 2.a and Fig. 2.b, we also noticed when Twitter users behave as receivers, the curves are smoother than those when they behave as information sources. This suggests that Twitter users can discover serendipity consistently, but they cannot provide serendipity at such a stable rate. The reason may be that Twitter users can discover information through a variety of means which add up to a serendipitous stream of information, e.g., hashtags (Kop 2012) and the indirect media exposure (An et al. 2011). As for outlier points in Fig. 2.a and Fig. 2.b, the reason is the intermittent problems in the phase of data collection.

In Fig. 2 and Fig. 3, the estimated proportions of unexpected, relevant, and serendipitous diffusion of information averaged along the time axis are also marked. For simplicity, we use AVG.Unexp, AVG.Rel and AVG.Serendipity as abbreviations. Again, the results are distinguished when the sampled users behave as receivers and sources. From the results, we can see that around 27% of information diffusion in Twitter are serendipitous, with the sampled users as either receivers or providers of information. The proportion of serendipitous information diffusion in Weibo is around 30%. Compared to Twitter users, Weibo users seemed to be a little better at providing and discovering serendipitous information, and more enthusiastic about posting and reposting relevant information. However, such observations have to be carefully confirmed with a larger sample from Weibo.

We believe the results provide persuasive evidence that serendipity has a strong presence in information diffusion in microblogging communities. More specifically, microblogging users are able to discover serendipitous content as a receiver, and provide serendipity as the source of information as well.

In the above analysis, there does exists a selection bias where we focus the diffusion of information through retweets only. We did this because in large-scale observational data, only retweeting is the reliable signal of a user actually "receiving" a tweet. The correction of this selection bias requests a reliable prediction model of information acquisition, which we leave as one of our future directions.

Effects of Serendipity

Upon proving the existence of serendipity in microblogging, we look into its effect in users' social statuses and activities. More specifically, we utilize statistical tests to investigate the effect of serendipity on users' future behaviors. To this end, we calculate the *change rate* of user attributes in a time span preceding the time period in which serendipity is identified. We ran the tests on Twitter data only, from which we can observe the complete set of user attributes. In Weibo, due to the limits of Weibo API, we were not able to collect all the required user attributes. Specifically, we use the held-out data from 11/1/2011 to 11/30/2011 (T_1) to measure the effect of serendipity. Recall that within the time span from 7/10/2011 to 10/30/2011 (T_0), we have conducted statistical tests and estimated the proportion of serendipity provided and received by each user.

The analysis of change rate is akin to the method of "difference in differences" (diff-in-diffs) in econometrics, which measures the effect of a treatment in a given time period (Meyer, Viscusi, and Durbin 1995; Card and Krueger 2000). For example, let T_0 be a time period before the treatment and T_1 be a time period after the treatment, the diff-in-diff of the attribute *friends* during the period $\Delta(T_0 \cup T_1)$ is denoted as $friends.\Delta$. Assume that for one user, 8/1 and 8/11 were the first and the last time we observed the user in T_0, and his number of friends was observed as 100 and 200 on the two days. Suppose 11/11 and 11/21 were the first and the last dates we observed him in T_1 and the number of his friends was observed as 300 and 450, respectively. Then, the daily change ratio of the number of friends in time periods T_0 and T_1 are 10 (i.e., $(200-100)/(11-1)$) and 15 (i.e., $(450-300)/(21-11)$). Thus $friends.\Delta$ is 0.5 (i.e., $(15-10)/10$).

Intuitively, if the diff-in-diff is greater than zero, there is likely to be a positive effect of the treatment on this attribute; otherwise the effect is negative. With this diff-in-diff analysis, we can test the effect of serendipity in user attributes that reflect the level of activities (the number of followers, friends, tweets, replies, retweets) and the social status (times of being retweeted, replied and listed). We computed the correlation between the proportion of serendipity a user provides or receives in T_0 and the diff-in-diff of each attribute in T_1. Specifically, we employed *Pearson's product moment correlation coefficient* as test statistic, to study whether the provision/exposure to serendipity have a positive effect on the social activities and status of users.

The results are given in Fig 4 and 5, where the sampled users served as the receivers and the sources, correspondingly. As the receiver, it shows that the more serendipitous information a user discovers, the more active (e.g., posting more tweets) and more sociable (e.g., retweeting and replying to others more frequently) he becomes. This is in line with findings in literature, where unexpected but relevant discovery brought more satisfaction and excitement to user experience, which leads to an enhanced user engagement. Meanwhile, as the source, the more serendipitous information a user provides, the more influential he becomes (e.g., being retweeted and replied to more frequently). The provision of serendipity also makes him more active and sociable.

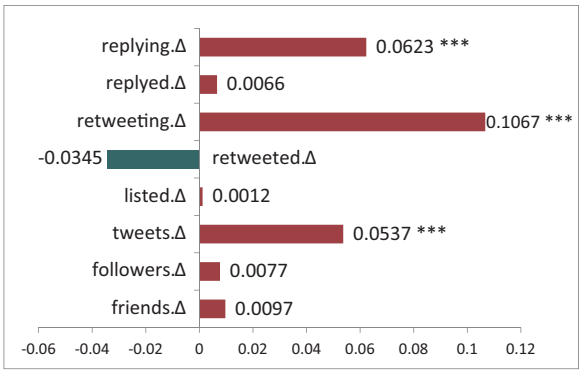


Figure 4: Discovering more serendipity increases the increase rates of social activities.

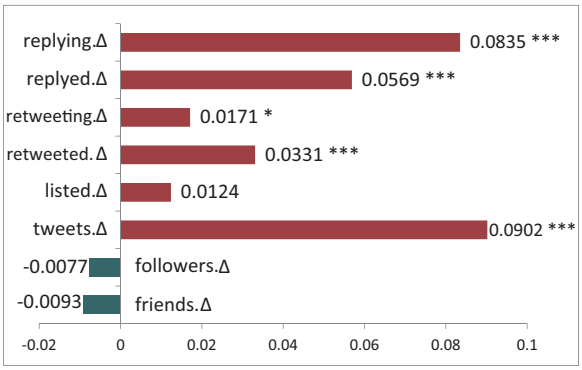


Figure 5: Providing more serendipity increases the increase rates of social activities and social influence.

Significance at the: *** 1%, ** 5%, or * 10% level

This can be explained by the fact that receivers who discover serendipity tend to interact more frequently with the sources, e.g., to express appreciation. Such activity increases the influence of the sources of serendipitous information, and in turn enhances the experience and engagement of the source users.

To sum up, we can draw the conclusion that serendipitous information diffusion in Twitter does have a positive effect on user activities and user engagement. The more serendipitous information diffused to a user, the more active and social he becomes. Meanwhile, the more serendipitous information diffused from a user, the more influential, active, and social he becomes.

Conclusion & Discussion

We presented the first quantitative study of serendipity with large-scale behavioral analysis, in the scenario of retweeting behaviors in microblogging communities. We formally defined serendipity as unexpected relevance, and proposed principled hypothesis tests to automatically identify unexpected, relevant, and serendipitous diffusion of information. We then showed the strong presence of serendipitous information diffused in microblogs, accounting for about 27% of retweets in Twitter and about 30% in Weibo. This surprisingly high ratio of serendipity is likely to contribute to the huge success of microblogging. Results of further statistical analysis demonstrated that both the discovery of serendipity and the provision of serendipity have a positive effect in the activeness, engagement, and influence of users. This suggests that the increase of serendipity may be an effective way to enhance user experience and engagement in microblogging.

We anticipate this study could provide insights to the design of information systems: we should be open to examining both aspects — the unexpectedness and the relevance. E.g., with the proposed method, in microblogs, we could identify and highlight serendipitous tweets in timelines, so that users can access serendipity more frequently; while in search engines, for serendipitous pages that are outside the reach of conventional IR models, we could rank them higher to bring more surprises.

The presented analysis of effects of serendipity on user engagement and social interaction is correlational rather than causal. Randomized controlled trials are needed to assess the causality of serendipity to user engagement. The causality test, the correction of selection bias, together with more sophisticated perceived models, will be our future directions.

Acknowledgement

This work is supported by the National Science Foundation under grant numbers IIS-1054199, IIS-0968489, and CCF-1048168, the National Natural Science Foundation of China (NSFC Grant No. 61272343) and the Doctoral Program of Higher Education of China (FSSP Grant No. 20120001110112).

References

Abbassi, Z.; Amer-Yahia, S.; Lakshmanan, L. V.; Vassilvitskii, S.; and Yu, C. 2009. Getting recommender systems to think outside the box. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, 285–288.

An, J.; Cha, M.; Gummadi, K.; and Crowcroft, J. 2011. Media landscape in Twitter: A world of new conventions and political diversity. In *Proceedings of International Conference on Weblogs and Social Media*, ICWSM'11.

Andel, P. V. 1994. Anatomy of the unsought finding. serendipity: Orgin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science* 45(2):631–648.

André, P.; Teevan, J.; and Dumais, S. T. 2009. From x-rays to silly putty via uranus: serendipity and its role in web search. In *CHI '09*, 2033–2036.

Bellotti, V.; Begole, B.; Chi, E. H.; Ducheneaut, N.; Fang, J.; Isaacs, E.; King, T.; Newman, M. W.; Partridge, K.; Price, B.; Rasmussen, P.; Roberts, M.; Schiano, D. J.; and Walendowski, A. 2008. Activity-based serendipitous recommendations with the magitti mobile leisure guide. In *CHI '08*, 1157–1166.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In

Proceedings of the 2010 43rd Hawaii International Conference on System Sciences, HICSS '10, 1–10.

Card, D., and Krueger, A. B. 2000. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review* 90(5):1397C1420.

Case, D. 2012. *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. Emerald, 3rd edition.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011. Political polarization on twitter. In *Proceedings of International Conference on Weblogs and Social Media*, ICWSM'11.

De Bruijn, O., and Spence, R. 2008. A new framework for theory-based interaction design applied to serendipitous information retrieval. *ACM Trans. Comput.-Hum. Interact.* 15(1):5:1–5:38.

Foster, A., and Ford, N. 2003. Serendipity and information seeking: an empirical study. *Journal of Documentation* 59(3):321–340.

Ge, M.; Delgado-Battenfeld, C.; and Jannach, D. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, 257–260.

Golin, M. 1957. Serendipity-big word in medical progress. *Journal of the American Medical Association* 165(16):2084–2087.

Granovetter, M. 1973. The strength of weak ties. *The American Journal of Sociology* 78(6):1360–1380.

Han, B.; Cook, P.; and Baldwin, T. 2013. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.* 4(1):5:1–5:27.

Herring, S. R.; Poon, C. M.; Balasi, G. A.; and Bailey, B. P. 2011. Tweetspiration: leveraging social media for design inspiration. In *CHI EA '11*, 2311–2316.

Huelsenbeck, J.; Hillis, D.; and Nielsen, R. 1996. A likelihood-ratio test of monophyly. *Systematic Biology* 45(4):546.

Jarvis, J. 2010. Serendipity is unexpected relevance. http://www.buzzmachine.com/2010/03/30/serendipity-is-unexpected-relevance/.

Kawamae, N. 2010. Serendipitous recommendations via innovators. In *SIGIR '10*, 218–225.

Kop, R. 2012. The unexpected connection: Serendipity and human mediation in networked learning. *Educational Technology & Society* 15(2):2–11.

Lin, J.; Snow, R.; and Morgan, W. 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *KDD '11*, 422–429.

Makri, S., and Blandford, A. 2011. What is serendipity? a workshop report. *Information Research* 16(3):491.

Manning, C.; Raghavan, P.; and Schutze, H. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press.

McCay-Peet, L., and Toms, E. G. 2011. The serendipity quotient. *Proceedings of the American Society for Information Science and Technology* 48(1):1–4.

Meyer, B. D.; Viscusi, W. K.; and Durbin, D. L. 1995. Workers' compensation and injury duration: Evidence from a natural experiment. *The American Economic Review* 85(3):322–340.

Murakami, T.; Mori, K.; and Orihara, R. 2008. Metrics for evaluating the serendipity of recommendation lists. In *Proceedings of the 2007 conference on New frontiers in artificial intelligence*, JSAI'07, 40–46.

Neyman, J., and Pearson, E. S. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philos. Trans. R. Soc.* 231:289–337.

Piao, S., and Whittle, J. 2011. A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections. In *SocialCom/PASSAT'11*, 910–915. IEEE.

Ponte, J. M., and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 275–281.

Posada, D., and Crandall, K. A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Reagans, R., and Zuckerman, E. W. 2001. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization Science* 12(4):502–517.

Roberts, R. M. 1989. *Serendipity: Accidental Discoveries in Science*. Wiley.

Toms, E. G. 2000. Understanding and facilitating the browsing of electronic text. *Int. J. Hum.-Comput. Stud.* 52:423–452.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* 9:60–62.

Woyke, E. 2011. Ebay tests serendipitous shopping. *Forbes* 188.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, 177–186.

Zhai, C., and Lafferty, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334–342. ACM.

Zhang, Y. C.; Ó Séaghdha, D.; Quercia, D.; and Jambor, T. 2012. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '12.

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, 22–32.