Mining Social Media with Social Theories: A Survey

Jiliang Tang
Computer Science & Eng
Arizona State University
Tempe, AZ, USA
Jiliang.Tang@asu.edu

Yi Chang
Yahoo!Labs
Yahoo!Inc
Sunnyvale,CA, USA
yichang@yahoo-inc.com

Huan Liu
Computer Science & Eng
Arizona State University
Tempe, AZ, USA
Huan.Liu@asu.edu

ABSTRACT

The increasing popularity of social media encourages more and more users to participate in various online activities and produces data in an unprecedented rate. Social media data is big, linked, noisy, highly unstructured and incomplete, and differs from data in traditional data mining, which cultivates a new research field - social media mining. Social theories from social sciences are helpful to explain social phenomena. The scale and properties of social media data are very different from these of data social sciences use to develop social theories. As a new type of social data, social media data has a fundamental question - *can we apply social theories to social media data?* Recent advances in computer science provide necessary computational tools and techniques for us to verify social theories on large-scale social media data. Social theories have been applied to mining social media. In this article, we review some key social theories in mining social media, their verification approaches, interesting findings, and state-of-the-art algorithms. We also discuss some future directions in this active area of mining social media with social theories.

1. INTRODUCTION

Social media greatly enables people to participate in online activities and shatters the barrier for online users to share and consume information in any place at any time. Social media users can be both passive content consumers and active content producers, and generate data at an unprecedented rate. The nature of social media determines that its data significantly differs from the data in traditional data mining. Social relations are pervasively available in social media data, and play important roles in social media such as mitigating information overload problem [38; 51] and promoting the information propagation process [4; 67].

Social media data is big, noisy, incomplete, highly unstructured and linked with social relations. These unique properties of social media data suggest that naively applying existing techniques may fail or lead to inappropriate understandings about the data. For example, social media data is linked via social relations and contradicts with the underlying independent and identically distributed (IID) assumption of the vast majority of existing techniques [23; 57]. This new type of data calls for novel data mining techniques for a better understanding from the computational perspective. The study and development of these new techniques are under the purview of social media mining, which is the process of representing, analyzing, and extracting actionable patterns from social media data [70].

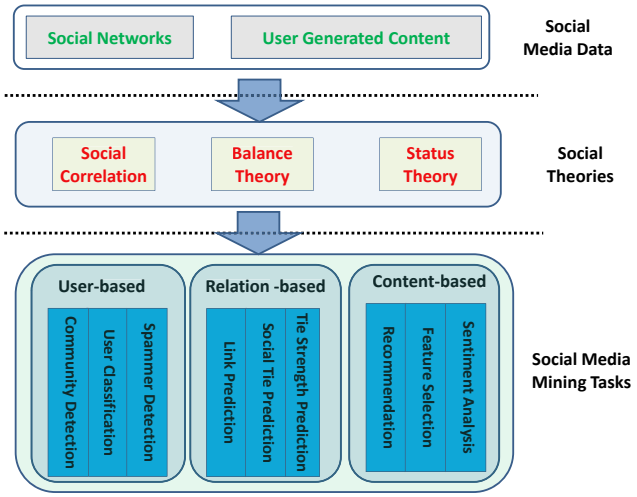

Figure 1: Social Theories in Social Media Mining.

There are many social theories developed from social sciences to explain various types of social phenomena. For example, the homophily theory [40] suggests how individuals connect to each other, while balance theory suggests that users in a social network tend to form into a balanced network structure [17]. The scale of the data social scientists employ to develop these social theories is very different from that of social media data. It is easy for social media data to include the actions and interactions of hundreds of millions of individuals in real time as well as over time. Therefore there is a fundamental question for this new type of social data - *can we apply some social theories to social media data?*. If we can apply social theories to social media data, social theories can help us understand social media data from a social perspective, and combining social theories with computational methods manifests a novel and effective perspective to mine social media data as shown in Figure 1. Social theories help bridge the gap from what we have (social media data) to what we want to understand social media data (social media mining).

Integrating social theories with computational models becomes an interesting direction in mining social media data and encourages a large body of literature in this line. The goal of this article is to provide a review of some key social

theories in mining social media data. The contributions and organization of this article are summarized as below:

- The social property of social media data determines that it differs from data in traditional data mining and social sciences. In Section 2, we provide an overview of the unique properties of social media data;
- An increasing number of social theories is verified in social media data. In Section 3, we focus on three key and widely used social theories with basic concepts, verification approaches and key findings;
- The fast growing interests and intensifying need to harness social media data make social media mining grow rapidly. Integrating social theories with computational methods becomes a principled way to mine social media data. In Section 4, we review the state-of-the-art algorithms that exploit social theories in mining social media, and summary feature engineering, constraint generating and objective defining as three ways to explain social theories for computational models.
- Social theories in mining social media data is still an active area of exploration and there could be more existing social theories to be employed or new social theories to be discovered from this new type of social media data. In Section 5, we discuss some open issues and possible research directions.

2. SOCIAL MEDIA DATA IS SOCIAL

Social relations are pervasively available and the social property of social media data determines that social media data is substantially different from data in traditional data mining and social sciences. In this section, we discuss some unique properties of social media data. Before details, we first introduce some notations used in this article. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_n\}$ and $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ be the set of n users and m items (or pieces of user generated content). We use $\mathbf{S} \in \mathbb{R}^{n\times n}$, $\mathbf{R} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times K}$ to denote user-user relation, user-content interaction and content-feature matrices where we extract a set of K features $\mathcal{F}$ to represent the content set $\mathcal{P}$.

Big: In social media, we have little data for each specific individual. However, the social property of social media data links individuals' data together, which provides a new type of big data. For example, more than 300 million tweets are sent to Twitter per day; more than 3,000 photos are uploaded to Flickr per minute, and more than 153 million blogs are posted per year.

Linked: The availability of social relations determines that social media data is inherently linked [52]. An illustration example is shown in Figure 2 where user generated content (or p_1 to p_8) are linked via social relations among users (u_1 to u_4). Linked social media data is patently not independent and identically distributed, which contradicts one of the most enduring and deeply buried assumptions of traditional data mining and machine learning methods [23; 57].

Noisy: A successful data mining exercise entails extensive data preprocessing and noisy removal as "garbage in and garbage out." However, social media data can contain a large portion of noisy data. Users in social media can be both passive content consumers and active content producers, causing the quality of user generated content to vary

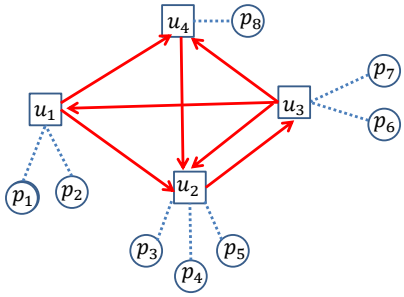

Figure 2: Linked Social Media Data.

drastically [1]. The noisy issues of social media data are not stop here. The social networks in social media are also noisy. First some social media users work as spammers to spread malicious or unwanted messages [47]. Second, the low cost of link formation leads to acquaintances and best friends mixed together [65].

Unstructured: User generated content in social media is often highly unstructured. Nowadays more and more users use their mobile devices to publish content such as updating statuses in Facebook, sending tweets in Twitter and commenting on posts, which results in (1) short texts and (2) typos and spacing errors occurring very frequently [25]. Free-from languages are widely adopted by social media users in the online communication such as ASCII art (e.g., :) and :( ) and abbreviations (e.g., h r u?) [46]. The short and highly unstructured social media data challenges the vast majority of existing techniques.

Incomplete: Users' attributes are predictable with their personal data [26]. To address such privacy concerns, social media services often allow their users to use their profile settings to mark their personal data such as demographic profiles, status updates, lists of friends, videos, photos, and interactions on posts, invisible to others. For example, a very small portion of Facebook users ($< 1\%$) make their personal data public available [41]. The available social media data could be incomplete and extremely sparse. For example, for social recommendation, more than 99% of entities in the user-content interaction matrix $\mathbf{R}$ are missed [51].

3. SOCIAL THEORIES

Social theories from social sciences are useful to explain various types of social phenomena. In social media, it is increasingly possible for us to observe social data from hundreds of millions of individuals. Given its large-scale size and social property, a natural question is - *can we apply social theories to social media data?*. More and more social theories have been proven to be applicable to social media data. In this section, we concentrate on three important social theories with basic concepts, ways to verify them and key findings.

3.1 Social Correlation Theory

Social correlation theory is one of the most important social theories and it suggests that there exist correlations between behaviors or attributes of adjacent users in a social network. Homophily, influence and confounding are three major social process to explain these correlations as shown in Figure 3.

- Homophily is to explain our tendency to connect to others that share certain similarity with us. For example, birds of a feather flock together.

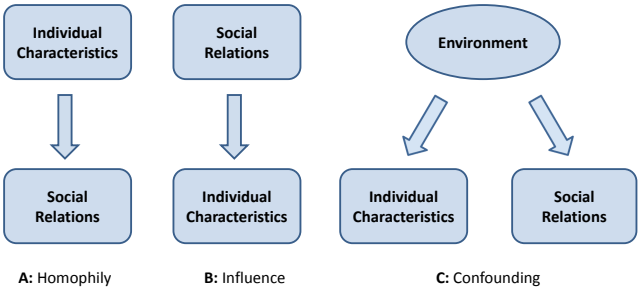

Figure 3: Major Social Forces of Social Correlation.

- Influence suggests that people tend to follow the behaviors of their friends and adjacent users are likely to exhibit similar behaviors. For example, if most of one's friends switch to a mobile phone company, he could be influenced by them and switch, too
- Confounding is a correlation between users that can also be forged due to external influences from environment. For example, two individuals living in the same city are more likely to become friends than two random individuals.

To help us verify the applicability of social correlation theory to social media data, essentially we need to answer the following question - *are users with social relations more similar than these without?* To answer this question, for each social relation from u_i to u_j, we calculate two similarities s_{ij} and r_{ik} where s_{ij} is the similarity between u_i and u_j, while r_{ik} is the similarity between u_i and a randomly chosen user u_k who does not connect to u_i. Let $\mathcal{S}$ be the set of s_{ij}s, which denotes the set of similarities of pairs of connected users. Let $\mathcal{R}$ be the set of r_{ik}s, which represents the set of similarities of pairs of randomly chosen users. We perform a t-test on $\mathcal{S}$ and $\mathcal{R}$. The null hypothesis is that similarities with social relations are no larger than these without, i.e., $H_0 : \mathcal{S} \leq \mathcal{R}$; the alternative hypothesis is that the similarities with social relations are larger than these without, i.e., $H_1 : \mathcal{S} > \mathcal{R}$. If there is strong evidence to reject the null hypothesis, we verify that social correlation theory is applicable to social media data.

Via above verification process, social correlation theory has been proven to be applicable to various social media sites. Twitter users with following relations are likely to share similar topics or opinions [63; 20]. Users in Epinions with trust relations are likely to rate same items with similar scores [49]. In [52], we shows that users in Digg and BlogCategory with social relations are likely to joint groups of similar interests. In location-based social networks such as Foursquare, users with social relations are likely to do checkins in the same locations [69; 14].

3.2 Balance Theory

In general, balance theory implies the intuition that "the friend of my friend is my friend" and "the enemy of my enemy is my friend" [17]. Basically, it considers the balance of signs on a triad involving three users in a social network with positive and negative links [28; 27]. We use s_{ij} to denote the sign of the relation between u_i and u_j where $s_{ij} = 1$ (or $s_{ij} = -1$) if we observe a positive relation (or a negative relation) between u_i and u_j. Balance theory suggests that

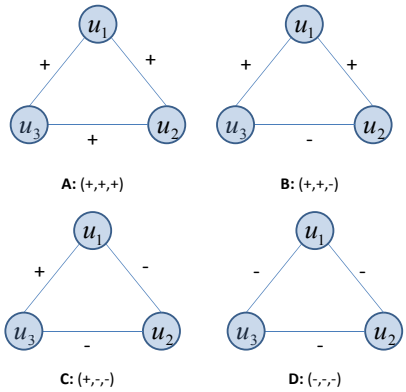

Figure 4: An Illustration of Balance Theory.

a triad $\langle u_i, u_j, u_k \rangle$ is balanced if

- $s_{ij} = 1$ and $s_{jk} = 1$, then $s_{ik} = 1$; or
- $s_{ij} = -1$ and $s_{jk} = -1$, then $s_{ik} = 1$.

For a triad $\langle u_i, u_j, u_k \rangle$, there are four possible sign combinations $\mathbf{A}(+,+,+)$, $\mathbf{B}(+,+,-)$ $\mathbf{C}(+,-,-)$ and $\mathbf{D}(-,-,-)$ as shown in Figure 4, while only $\mathbf{A}(+,+,+)$ and $\mathbf{C}(+,-,-)$ are balanced. The way to verify balance theory is straightforward. We examine all these triads $\langle u_i, u_j, u_k \rangle$ and then to check the ratio of $\mathbf{A}(+,+,+)$ and $\mathbf{C}(+,-,-)$ among all four possible sign combinations. A high ratio suggests that balance theory is applicable to social media data

We check the distributions of four possible sign combinations on the three widely used social media datasets (i.e., Epinions, Slashdot and Wikipedia) with signed networks in [27]. The ratios of $\mathbf{A}(+,+,+)$ and $\mathbf{C}(+,-,-)$ among all four possible sign combinations are 0.941, 0.912, and 0.909 in Epinions, Slashdot and Wikipedia, respectively. More than 90% of triads are balanced. Similar observations on other social media datasets are reported by [68; 56]. Note that balance theory is developed for undirected social networks and we usually ignore their directions when applying balance theory to directed social networks [27].

3.3 Status Theory

Different from balance theory, status theory is developed for directed social networks [28]. Social status refers to the position or rank of a user in a social community, and represents the degree of honor or prestige attached to the position of each individual. In status theory, a positive link from u_i to u_j indicates that u_i has a higher status than u_j; while a negative link from u_i to u_j indicates that u_i has a lower status than u_j. For a triad, status theory suggests that if we take each negative relation, reverse its direction, and flip its sign to positive, then the resulting triangle (with all positive edge signs) should be acyclic.

In [28], contextualized links are introduced to verify status theory. A contextualized link is defined to be a triple $\langle u_i, u_j, u_k \rangle$ with the property that a link forms from u_i to u_j after each of u_i and u_j already has a link either to or from u_k. The link between u_k and u_i can go in either direction and have either sign yielding four possibilities, and similarly for the link between u_k and u_j; hence overall there are $4 \times 4 = 16$ different types of contextualized links. Figure 5 demonstrates 4 of 16 types of contextualized links where ($\mathbf{A}$)

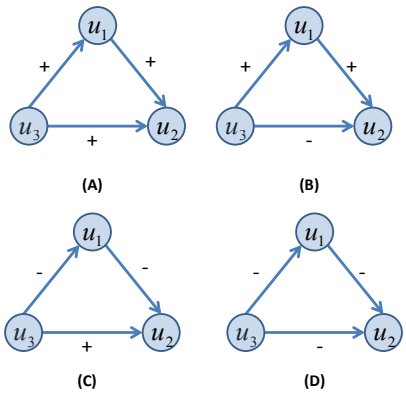

Figure 5: An Illustration of Four Out of Sixteen Types of Contextualized Links for Status Theory. Note that "+"(or "-") denotes the target node has a higher (or lower) status than the source node.

and (**D**) satisfy the status theory, while (**B**) and (**C**) do not satisfy the status theory. For each of these types of contextualized links, we can count frequencies of positive versus negative links for the links from u_i to u_j and then calculate the ratio of contextualized links satisfying status theory.
In [53], it is reported that 99% of triads in the Enron email social network and the advisor-advisee social network satisfy status theory. Similar patterns are observed on Epinions and Wikipedia datasets in [28].

3.4 Discussion

The scale and properties of social media data substantially differ from these of data used by social sciences to develop social theories. Since social media data is a new type of social data, it is possible to apply some social theories to explain phenomena in social media data. The verification of social theories in social media data not only paves a way for us to understand social media data from a social perspective but also suggests that it is highly possible to facilitate social media mining tasks by integrating social theories with computational methods.

4. SOCIAL THEORIES IN SOCIAL MEDIA MINING TASKS

Social media mining is an emerging discipline under the umbrella of data mining and grows rapidly in recent years [70]. The verification of some social theories in social media data suggests that we should put "social" in social media mining and encourages a large body of literature to model and exploit social theories to advance social media mining tasks. In general, there are three types of objects in social media data - users, social relations and user generated content, which allows us to roughly classify social media mining tasks to three groups based on the mining objects - user-based tasks, relation-based tasks and content-based tasks. Next we elaborate each group with representative tasks with their definitions, challenges and the start-of-the-art algorithms to apply social theories to these specific tasks.

4.1 Social Theories in User-Related Tasks

For individuals, a better understanding of their social networks can help them share and collect reliable information more effective and efficient. For social media service providers, a better understanding of their customers can help them provide better services. User-related tasks provide necessary and effective means to understand social media users. In this subsection, we review social theories in some key user-related tasks.

4.1.1 Community Detection

Communities in social media can be explicit such as Yahoo! Groups. However, in many social media sites, communities are implicit and their members are obscure to social media users. Community detection is proposed to find these implicit communities in social media by identifying groups of users that are more densely connected to each other than to the rest of the network [55]. Detecting implicit communities can benefits many social media mining tasks such as social targeting and personalization. The major difference between clustering in data mining and community detection is that in community detection, individuals are connected to others in social networks; while in clustering, data points are not embedded in a network and they are assume to be independent and identically distributed. Formally, for a social network $G(\mathcal{U}, \mathbf{S})$, community detection is to find a set of communities $\mathcal{C}$ where users are more densely connected within a community than to the rest of users.
Homophily suggests that similar users are likely to be linked, and influence indicates that linked users will influence each other and become more similar. The suggestions from social correlation theory in creating new ties based on the similarity gives rise to macro patterns of associations, also known as communities [7]. Two users in the same community have higher similarity [44]. The modularity maximization method is to maximize the sum of the actual number of social relations between two users minus expected number of social relations between them since two users in the same community should have a higher probability to establish a relation than two randomly chosen users [43]. Wang et al. [60] find that users within the community are likely to share similar tags in social tagging systems and they take advantage of the bipartite network between users and tags in social tagging systems to discover these overlapping communities. In [66], a density-based framework is proposed with the intuition that users in the same community should interact more frequently with each other.
Recently applying balance theory to detect communities from signed networks has attracted increasing attention. In [11], a generalized balance theory is proposed where a network is k-balanced iff users can be partitioned in to k-subsets such that positive links lie within the sets and the negative links between them. Balance theory suggests that the assignment of users related by negative links should be done the opposite way of positive links, with negative links sparse within and more dense between communities therefore the potts model is extended to incorporate both positive and negative links to detect communities in signed network [58]. In [2], a two-objective approach is proposed for community detection in signed networks based on balance theory. One is that the partitioning should have dense positive intra-connections and sparse negative interconnections, and the other is that it should have as few as possible negative intra-connections and positive inter-connections.

4.1.2 User Classification

Due to privacy concerns, social media users tend to hide their profiles. For social media service providers, users' profile information is useful for them to customize their services to the users in many ways such as friend and content recommendations and personalized search. More they know about users and their preferences, better they can serve them. Given a social network and some user information (attributes, preferences or behaviors), user classification is designed to infer the information of other users in the same network [15]. In the user classification problem, users in $\mathcal{U}$ are partially labeled as $\mathcal{U} = [\mathcal{U}^L, \mathbf{U}^U]$ where $\mathcal{U}^L$ and $\mathbf{U}^U$ are the sets of labeled and unlabeled users, respectively. Formally the task of user classification is to label users from a finite set of categorical values in $\mathbf{U}^U$ with the social network $G(\mathcal{U}, \mathbf{S})$ and $\mathcal{U}^L$.

Social correlation theory suggests that the labels of linked users should be correlated, which is the major reason why researchers believe that the labels of $\mathcal{U}^L$ can be predicted with the network structure and the partially labeled users [15]. Social correlation theory is the underlying assumption of most of existing user classification methods, which design algorithms for collective classification. A typical user classification algorithm includes parts of the three components [37]:

- A local classifier - it is used for initial label assignment;
- A relational classifier - it learns a classifier from the labels of its neighbors to the label of one user suggested by social correlation theory; and
- Collective classification - it applies relational classifier to each node iteratively until the inconsistency between neighboring labels is minimized.

In [36], a weighted-vote relational neighborhood classifier wvRN is introduced for user classification. wvRN is like a lazy learner and estimates the labels of users as the weighted mean of their neighbors. In [34], the proposed framework first creates relational features of one user by aggregating the label information of its neighbors and then a relational classifier can be constructed based on labeled data. Neville and Jensen in [42] propose to use clustering algorithms to find out the cluster memberships of each user first, and then fix the latent group variables for later inference. Xiang et al. [64] propose a novel latent relational model based on copulas. It can make predictions in a discrete label space while ensuring identical marginals and at the same time incorporating some desirable properties of modeling relational dependencies in a continuous space. A community-based framework is proposed in [54]. It first extracts overlapping communities based on social network structure, then uses communities as features to represent users and finally a traditional classifier such as SVM is trained to assign labels for unlabeled users in the same network.

4.1.3 Social Spammer Detection

Social media has become an important and efficient way to disseminate information. Given its popularity and ubiquity, social spammers create many fake accounts and send out unsolicited commercial content [62]. Social spammers have become rampant and the volume of spam has increased dramatically. For example, 83% of the users of social networks have received at least one unwanted friend request or message [47]. This not only causes misuse of communication bandwidth, storage space and computational power, but also wastes users' time and violates their privacy rights. Therefore developing effective social spammer detection techniques is critically important in improving user experience and positively affecting the overall value of social media services [47]. Given a social network $G(\mathcal{U}, \mathbf{S})$, social spammer detection is to find a set of spammers $\mathcal{U}^S$ from $\mathcal{U}$ with $\mathcal{U}^S \subset \mathcal{U}$.

Based on social correlation theory, there are two observations for normal users and spammers [73]. First normal users perform similarly with their neighbors. Second, spammers perform differently from their neighbors since most of their neighbors are normal users. Therefore a social regularization term is proposed under the matrix factorization framework to model these observations where two connected normal users should be close in the latent space since they share similar interests and may perform similar social activities, while spammers should be far away from their neighbors in the latent space. In Twitter, users have directed following relations and spammers can easily follow a large number of normal users within a short time. In [19], we divide user-user following relations in Twitter into four types - [spammer, spammer], [normal, normal], [normal, spammer], and [spammer, normal]. Since the fourth relation can be intentionally faked by spammers, we only consider the first three types of relations. Specifically we introduce a graph regularization term to model social correlation theory in the directed social relations, which is integrated into the standard Lasso formulation to train a linear classification for social spammer detection. Spammers and normal users have very different social behaviors. Normal users are likely to form a group with other normal users, while spammers are likely to from spammer groups [29]. In [6], the authors incorporate community-based features of users with basic topological features to improve spammer classifiers. It first finds overlapping community structure of users and then extracts features based on these communities such as the features which express the role of a user in the community structure like a boundary node or a core node and the number of communities it belongs to.

4.2 Social Theories in Relation-Related Tasks

A social network is usually represented by a binary adjacent matrix. First the matrix is extremely sparse since there are many pairs of users with missing relations. Second, social networks in social media are more complicated. For example, strengths of relations might be heterogeneous such as acquaintances and best friends, while a social network may a composite of various types of relations such as family, classmates and colleagues. Relation-Related tasks focus on mining relations among users and aim to reveal a fine-grained and comprehensive view of social relations. Signed networks arise in social network with various ways when users can implicitly or explicitly tag their relationship with other users as positive or negative. In this section, we review social theories in some key relation-related tasks on signed and unsigned networks.

4.2.1 Link Prediction

It is critical for social media sites to provide services to encourage more user interactions with better experience such as expanding one's social network. One effective way is to automatically recommend connections since it is hard for users to figure out who is available on social media sites.

Most social media sites provide friend recommendation services to their customers such as Facebook, Twitter and LinkedIn. The essential problem of friend recommendation is known as link prediction [30]. When there is no relation between u_i and u_j, $\mathbf{S}_{ij} = 0$. The task of link prediction is to predict which pairs of users u_i and u_j without relations $\mathbf{S}_{ij} = 0$ are likely to get connected given a social network $G(\mathcal{U}, \mathbf{S})$.

Unsigned Networks : Homophily in social correlation theory suggests that similar users are likely to establish social relations. In [30], various similarity measurements such as common neighbors based on the network structure are reviewed for link prediction. One challenging problem in link prediction is the sparsity problem - some users may have very few or even no links. In [49], a low-rank matrix factorization framework with homophily effect hTrust is proposed to predict trust relations. Homophily coefficients are defined to measure the strength of homophily among users. The stronger homophily between two users is, the smaller distance between them in the latent space is. Homophily regularization is then defined to model homophily effect by controlling users' distances in the latent space with the help of homophily coefficients. Through homophily regularization, trust relations can be suggested to users with few or even no relations and mitigate the sparsity problem in link prediction. The confounding effect in social correlation theory suggests that people who share high degree of overlap in their trajectories are expected to have a better likelihood of forming new links. In [59], the effect of confounding is investigated for link prediction. Specifically, it leverages mobility information to extract features which can capture some degree of closeness in physical world between two individuals. Status theory suggests new links are more likely to be attached from users with low statuses to these with high statues and the preferential attachment models are widely used to predict link prediction based on status measures such as the degree of nodes and PageRank [5].

Signed Networks : In [27], local-topology-based features (or 16 triad types) based on balance theory and status theory are extracted to improve the performance of a logistic regression classifier in signed relation prediction. In [13], the authors use a probabilistic treatment of trust combined with a modified spring-embedded layout algorithm to classify a relation based on balance theory. Instead of having all users repel, the model adds a repelling force only between users connected with a negative relation to capture balance theory. For example, one is friends with an enemy of the other; the forces will push them in different locations. In [10], the authors show how any quantitative measure of social imbalance in a network can be used to derive a link prediction algorithm and extend the approach in [27] by presenting a supervised machine learning based link prediction method that uses features derived from longer cycles in the network. The motivation to derive features from longer cycles is that higher order cycles in a signed network yield a "measure of imbalance" suggested the balance theory. In [18], it shows that the notion of weak structural balance in signed networks naturally leads to a global low-rank model for the network. Under such a model, the sign inference problem can be formulated as a low-rank matrix completion problem.

4.2.2 Social Tie Prediction

Social networks in social media can be a composite of various types of relations. For example, the relation types in Facebook could be family, colleagues, classmates and friends. However, in most online networks such as Facebook, Twitter and LinkedIn, such type information is usually unavailable [56]. Different types of relations may influence people in different ways. For example, one user's work style may be mainly influenced by her/his colleagues; while the daily life habits may be strongly affected by her/his family. It is necessary and important to reveal these different types of social relations therefore we ask whether we can automatically infer the types of social relations for social networks in social media. A novel task of social tie prediction is designed to answer the above question, which aims to predict the type of a given social relation. A non-zero value of $\mathbf{S}_{ij}$ suggests that there is a connection between u_i and u_j. Formally social tie prediction is to predict the type of a social relation between u_i and u_j with $\mathbf{S}_{ij} \neq 0$ from a finite set of categorical types such as { family, classmates, colleagues and friends}.

In [53], a framework is proposed to classify the type of social relationships by learning across heterogeneous networks. The framework incorporates social theories such as balance theory and status theory into a factor graph model, which effectively improves the accuracy of inferring the type of social relationships in a target network by borrowing knowledge from a different source network. Balance theory and status theory should be general over different types of networks. To learn knowledge from the source network to the target network, transfer features are extracted based on balance theory and status theory, which are shared by different types of networks. In particular, from social balance, the paper extracts triad based features to denote the proportion of different balanced triangles in a network; and from status theory, it defines features over triads to respectively represent the probabilities of the seven most frequent formations of triads. Different from [53], approaches are suggested by [68] to model balance theory and status theory mathematically. To model the balance theory, it introduces an one-dimensional latent factor β_i for each user u_i and defines the sign between u_i and u_j as $s_{ij} = \beta_i \beta_j$. To model status theory, it introduces a global user-independent parameter η to capture the partial ordering of users. η maps the latent user profile of u_i γ_i to a scalar quantity $\ell_i = \eta \gamma_i$, which reflects the corresponding user u_i's social status. According to status theory, it characterizes social ties from u_i to u_j by modeling the relative status difference between them as $\ell_{ij} = \ell_i - \ell_j$

4.2.3 Tie Strength Prediction

Social media users can have hundreds of social relations. However, a recent study shows that Twitter users have a very small number of friends compared to the number of followers and followees they declare [21]. The low cost of link formation in social media can lead to networks with heterogeneous relationship strengths (e.g., acquaintances and best friends mixed together) [65]. Pairs of users with strong strengths are likely to share greater similarity than those with weak strengths; therefore a better understanding of strengths of social relations can help social media sites serve their customers well such as better recommendations and more effective friend management tools, which arises the problem of tie strength prediction. In the binary relation presentation, once there is a connection between u_i and u_j, $\mathbf{S}_{ij} = 1$. The task of tie strength prediction is to predict a connection strength between 0 and 1 for u_i and u_j with

$\mathbf{S}_{ij} = 1$. After tie strength prediction, the binary relation representation matrix $\mathbf{S}_{ij} \in \{0, 1\}$ will be converted into a continuous valued relation representation matrix $\mathbf{S}_{ij} \in [0, 1]$. In [24], guided by social correlation theory, four different categories of features, i.e., attribute similarity, topological connectivity, transactional connectivity, and network transactional connectivity, are extracted from sources including friendship links, profile information, wall postings, picture postings, and group memberships. Then various classifiers are trained to predict link strength from transactional information based on these extracted features. A unsupervised latent variable model is proposed to predict tie strength in online social network [65] with user profiles and interactions. One key underlying assumption of the proposed model is social correlation theory. Homophily in social correlation theory postulates that users tend to form ties with other people who have similar characteristics, and it is likely that the stronger the tie, the higher the similarity. Thus the proposed framework models the tie strength as homophily effect of nodal profile similarities. The relationship strength directly influences the nature and frequency of online interactions between a pair of users. The stronger the relationship, the higher likelihood that a certain type of interaction between the pair of users. Therefore the propose framework models the relationship strength as the hidden cause of influence among users.

4.3 Social Theories in Content-Related Tasks

Numerous techniques are developed for various content mining tasks such as classification and clustering in the last decade. User generated content in social media is usually linked, noisy, highly unstructured and incomplete, which determines that existing techniques become difficult when applying these mining tasks on user generated content in social media. Before the popularity of social media, researchers have already noticed that exploiting link information can improve content classification [72] and clustering [32]. The popularity of social media makes social relations pervasively available, which encourages the exploitation of social relations in more and more mining tasks. Social theories can help us understand social relations better and in this subsection, we review how social theories help some representative content-related tasks.

4.3.1 Social Recommendation

The pervasive use of social media generates massive data in an unprecedented rate and the information overload problem becomes increasingly serve for social media users. Recommendation has been proven to be effective in mitigating the information overload problem and presents its significance to improve the quality of user experience, and to positively impact the success of social media. Users in the physical world are likely to seek suggestions from their friends before making a purchase decision and users' friends consistently provide good recommendations [45], we have similar observations in the online worlds. For example, 66% of people on social sites have asked friends or followers to help them make a decision and 88% of links that 14-24 year olds clicked were sent to them by a friend and 78% of consumers trust peer recommendations over ads and Google SERPs[1]. These intuitions motive a new research direction of recommendation social recommendation, which aims to take advantage of social relations to improve the performance of recommendation. Formally, a social recommender system is to predict missing values in the user-content interaction matrix $\mathbf{R}$ based on information from the user-user relation matrix $\mathbf{S}$ and the observed values in $\mathbf{R}$ [51].

The major reason why people believe that social relations are helpful to improve recommendation performance is evidence from social correlation theory, which suggests that a user's preference is similar to or influenced by their directly connected friends [51]. Therefore social media users rarely make decisions independently and usually seek advice from their friends before making purchase decisions. Social relations may provide both similar and familiar evidence for users, MoleTrust uses socially connected users to replace similar users in traditional user-based collaborative filtering method for recommendation in [39]. Social correlation theory indicates that a user's preference should be similar to her/his social network. Ensemble methods predict a missing value for a given users as a linear combination of ratings from the user and her/his social network based on traditional matrix factorization CF method with the intuition that users and their social networks should have similar ratings on the same items [50]. While regularization methods add regularization terms to force the preference of a user close to that of users in her/his social network under the matrix factorization CF method. For example, SocialMF defines a regularization term to force the preference of a user to be close to the average preference of the user's social network [22], and SoReg uses social regularization to force the preferences of two connected users close [35].

4.3.2 Feature Selection

One characteristic of user generated content in social media is high-dimensional such as there are tens of thousands of terms in tweets or pixels for photos in Flickr. Traditional data mining tasks such as classification and clustering may fail due to the curse of dimensionality. Feature selection has been proven to be an effective way to handle high-dimensional data for efficient data mining [31]. As mentioned above, user generated content is linked due to the availability of social relations and poses challenges to traditional feature selection algorithms which are typically designed for IID data. The formal definition of feature selection for user generated content in social media is stated as [52] - we aim to develop a selector which selects a subset of most relevant features from $\mathcal{F}$ on the content-feature matrix $\mathbf{C}$ with its social context $\mathbf{S}$ and $\mathbf{R}$.

LinkedFS is proposed as a feature selection framework for user generated content with social context based on social correlation theory in [52]. Four types of relations, i.e., co-Post, coFollowing, coFollowed and Following, are extracted from social context $\mathbf{S}$ and $\mathbf{R}$ of user generated content $\mathbf{C}$. Social correlation theory suggests that linked users are likely to share similar topics. Based on social correlation theory, LinkedFS turns these four types of relations to four corresponding hypotheses that can affect feature selection with linked data. For example, following hypothesis assumes that one user u_i follows another user u_j because u_i share u_j's interests, and their user generated content is more likely similar in terms of topics; hence LinkedFS models following relations mathematically by forcing topics of two users with

[1]http://www.firebellymarketing.com/2009/12/social-search-statistics-fromses-chicago.html

following relations close to each other. LinkedFS jointly incorporates group Lasso with the regularization term to model each type of relations for feature selection.

4.3.3 Sentiment Analysis

Nowadays social media services such as Twitter and Facebook are increasingly used by online users to share and exchange opinions, providing rich resources to understand public opinions. For example, in [3], a simple model exploiting Twitter sentiment and content outperforms market-based predictors in terms of forecasting box-office revenues for movies; public mood as measured from a large-scale collection of tweets obtains an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA [8]. Therefore sentiment analysis for such opinion-rich social media data has attracted increasing attention in recent years [46; 20]. Formally sentiment analysis for user generated content with social relations is to obtain a predictor from the content-feature matrix $\mathbf{C}$ with its social context $\mathbf{S}$ and $\mathbf{R}$, which can automatically label the sentiment polarity of an unseen post.

Social correlation theory indicates that sentiments of two linked users are likely to be similar. In [48], graphical models are proposed to incorporate social network information to improve user-level sentiment classification of different topics based on two observations - (1) user pairs in which at least one party links to the other are more likely to hold the same sentiment, and (2) two users with the same sentiment are more likely to have at least one link to the other than two users with different sentiment. Social correlation theory suggests that social relations are kinds of sentiment correlations. In [46], the authors propagate sentiment labels of tweets via user-user social relations $\mathbf{S}$ and user-tweet relations $\mathbf{R}$ to assign sentiment labels to unlabeled tweets. In [20], tweet-tweet correlation network are built from $\mathbf{S}$ and $\mathbf{R}$ based on social correlation theory. For example, tweets from users with following relations should be correlated as suggested by social correlation theory. Two tweets linked in the tweet-tweet correlation network are likely to share similar sentiments; hence the proposed framework SANT adds a graph regularization term in the Lasso classifier to force the sentiments of two correlated tweets close to each other.

4.4 Discussion

In reviewing state-of-the-art algorithms that exploit social theories in mining social media, we understand that they aim to find mathematical explanations of social theories for computational models. We notice that algorithms share similar ways in applying social theories such as feature engineering, constraint generating and objective defining.

- *Feature Engineering:* It uses social theories to extract features for computational models. For example, in link prediction, confounding effect in social correlation theory suggests that people who are physically close have a better likelihood of forming new links and new features from users' mobility information are extracted in [59] to improve link predilection; while triad features based on status theory are extracted as transfer features to infer social ties by transferring knowledge from the source network to the target network [53].

- *Constraint Generating:* It generates constraints from social theories for computational models. Regularization is one of the most popular ways to implement constraint generating. For example, SocialMF in social recommendation adds a social regularization term to force the performance of a user close to that of her/his social network to capture social correlation theories [22]; and hTrust adds a homophily regularization term to capture homophily effect and mitigate the sparsity problem in link prediction [49].

- *Objective Defining:* It uses social theories to define the objectives of the computational models. For example, two objectives are defined from balance theory to detect communities in signed networks [2]; and the user classification task is to make the labels of a user similar to these of her/his social network [15].

Social Media Mining Tasks		Feature Engineering	Constraint Generating	Objective Defining
User Related	Community Detection			[42],[59],[65],[57],[2]
	User Classification			[35],[33],[41],[63],[53]
	Spammer Detection	[28],[6]	[72],[18]	
Relation Related	Link Prediction	[29],[58],[26],[10]	[48]	[5],[12],[17]
	Social Tie Prediction	[52]		[67]
	Tie Strength Prediction	[23]		[64]
Content Related	Recommendation		[21],[23]	[38]
	Feature Selection		[51]	
	Sentiment Analysis		[47],[19]	[45]

Figure 6: Social Theories in Social Media Mining.

Instead of brute-force search, social theories can guide us to extract relevant features via feature engineering, to generate constraints via constraint generating, and to define objectives via objective defining for computational models. The algorithms reviewed earlier that exploit social theories in various social media mining tasks are summarized in Figure 6. We notice that for the same task, social theories can be exploited in different ways. For example, for link prediction, social theories are explained via feature engineering, constraint generating and objective defining.

5. OPEN ISSUES AND FUTURE RESEARCH DIRECTIONS

5.1 More "Social" in Mining Social Media Data

Some social theories have been proven to be applicable to social media data, which encourages us to put "social" in social media mining. Integrating some social theories with computational models advances various social media mining tasks and has attracted increasing attention. The exciting progress not only proves that the direction of integrating social theories in mining social media data is appealing but also suggests that we should put more "social" in social media mining. In this article, we review the state-of-the-art algorithms that employ social correlation theory, balance theory and status theory in various social media mining tasks. These theories are just illustrative examples and there could be more social theories to be applicable and employed such as small world theory [74] as shown in recent efforts to investigate and verify more social theories for social media data. Some of these efforts have already made initial progress such as structural hole theory [9] and weak tie theory [16]. A person is said to span a structural hole in a social network if he or she is linked to people in parts of the network that

are otherwise not well connected to one another [9]. Tang et al. [56] employ structural hole theory in the problem of social tie prediction; while Lou and Tang confirm the importance of structural hole in information diffusion with social media data, and show that mining structural hole can benefit various social media mining tasks such as community detection and link prediction [33]. Weak tie theory suggests that more novel information flows to individuals through weak rather than strong ties [16]. Recently researchers find that weak ties of a user are helpful to predict the preference of the user for user classification [54] and social recommendation [71].

5.2 New Social Theories

No doubt that social media data is a new type of social data and is much more complicated than the data social sciences use to study social theories. It is highly possible that new social theories can be discovered from social media data to make meaningful progress on important problems in social media mining, however, that progress requires serious engagement of both computer scientists and social scientists [61]. Data availability is still a challenging problem for social scientists. The data required to address many problems of interest to social scientists remain difficult to assemble and it has been impossible to collect observational data on the scale of hundreds of millions, or even tens of thousands, of individuals [61]. Social media provides a virtual world for users' online activities and makes it possible for social scientists to observe social behavior and interaction data of hundreds of millions of users. However social media data is too big to be directly handled by social scientists. On the other hand, computer scientists can employ data mining and machine learning techniques to handle big social media data; but, we lack necessary theories to help us understand social media data better. For example, without a better understanding of social media data, computer scientists may waste a lot of time in feature engineering, which is the key to the success of many real-world applications [12]. Therefore engagement of both computer scientists and social scientists in social media data is truly mutually beneficial. Computer scientists can take advantage of social theories to mine social media data and provide computational tools that are of great potential benefit to social scientists; while social scientists can make use of computational tools to handle social media data and develop new social theories to help computer scientists provide better computational tools.

6. CONCLUSION

The social nature of social media data calls for new techniques and tools and cultivates a new field - social media mining. Social theories from social sciences have been proven to be applicable to mining social media. Integrating social theories with computational models is becoming an interesting way in mining social media data and makes exciting progress in various social media mining tasks. In this article, we review three key social theories, i.e., social correlation theory, balance theory and status theory, in mining social media data. In detail, we introduce basic concepts, verification methods, interesting findings and the state-of-the-art algorithms to exploit these social theories in social media mining tasks, which can be categorized to feature engineering, constraint generating and objective defining. As future directions, more existing social theories could be employed or new social theories could be discovered to advance social media mining.

Acknowledgments

This work is, in part, supported by NSF (#IIS-1217466), ARO (#025071), ONR (N000141410095), Minerva Grant (N000141310835) and a research fund from Yahoo Faculty Research and Engagement Program.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, 2008.

[2] A. Amelio and C. Pizzuti. Community mining in signed networks: a multiobjective approach. In *ASONAM*, 2013.

[3] S. Asur and B. A. Huberman. Predicting the future with social media. In *WI-IAT*, 2010.

[4] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

[5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 1999.

[6] S. Y. Bhat and M. Abulaish. Community-based features for identifying spammers in online social networks. In *ASONAM*, pages 100–107. ACM, 2013.

[7] H. Bisgin, N. Agarwal, and X. Xu. Investigating homophily in online social networks. In *WI-IAT*, 2010.

[8] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[9] R. S. Burt. *Structural holes: The social structure of competition*. Harvard university press, 2009.

[10] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon. Exploiting longer cycles for link prediction in signed networks. In *CIKM*, 2011.

[11] J. A. Davis. Clustering and structural balance in graphs. *Human relations*, 1967.

[12] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 2012.

[13] T. DuBois, J. Golbeck, and A. Srinivasan. Predicting trust and distrust in social networks. In *socialcom*, 2011.

[14] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[15] L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 2005.

[16] M. Granovetter. The strength of weak ties. *JSTOR*, 1973.

[17] F. Heider. Attitudes and cognitive organization. *The Journal of psychology*, 1946.

[18] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon. Low rank modeling of signed networks. In *KDD*, 2012.

[19] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *IJCAI*, 2013.

[20] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *WSDM*, 2013.

[21] B. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 2008.

[22] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Recsys*, 2010.

[23] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *ICML*, 2002.

[24] I. Kahanda and J. Neville. Using transactional information to predict link strength in online social networks. In *ICWSM*, 2009.

[25] D. Kim, D. Kim, E. Hwang, and S. Rho. Twittertrends: a spatio-temporal trend detection and related keywords recommendation scheme. *Multimedia Systems*,2014.

[26] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.

[27] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

[28] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI*, 2010.

[29] F. Li and M.-H. Hsieh. An empirical study of clustering behavior of spammers and group-based anti-spam strategies. In *CEAS*, 2006.

[30] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 2007.

[31] H. Liu and H. Motoda. *Computational methods of feature selection*. CRC Press, 2007.

[32] B. Long, Z. M. Zhang, X. Wu, and P. S. Yu. Spectral clustering for multi-type relational data. In *ICML*, 2006.

[33] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW*, 2013.

[34] Q. Lu and L. Getoor. Link-based classification. In *ICML*, 2003.

[35] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *WSDM*, 2011.

[36] S. A. Macskassy and F. Provost. A simple relational classifier. In *MRDM*, 2003.

[37] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 2007.

[38] P. Massa. A survey of trust use and modeling in real online systems. *Trust in E-services: Technologies, Practices and Challenges*, 2007.

[39] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. In *CoopIS, DOA, and ODBASE*, 2004.

[40] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.

[41] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, 2010.

[42] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *MRDM*, 2005.

[43] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *PRE*, 69(2):026113, 2004.

[44] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *DMKD*, 2012.

[45] R. R. Sinha and K. Swearingen. Comparing recommendations made by online systems and friends. In *DELOS*, 2001.

[46] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *ULNLP*, 2011.

[47] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *ACSAC*, 2010.

[48] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *KDD*, 2011.

[49] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *WSDM*, 2013.

[50] J. Tang, H. Gao, and H. Liu. mtrust: discerning multi-faceted trust in a connected world. In *WSDM*, 2012.

[51] J. Tang, X. Hu, and H. Liu. Social recommendation: a review. *SNAM*, 2013.

[52] J. Tang and H. Liu. Feature selection with linked data in social media. In *SDM*, 2012.

[53] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM*, 2012.

[54] L. Tang and H. Liu. Relational learning via latent social dimensions. In *KDD*, 2009.

[55] L. Tang and H. Liu. Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2010.

[56] W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *PKDD*, 2011.

[57] B. Taskar, P. Abbeel, M.-F. Wong, and D. Koller. Label and link prediction in relational data. In *SRL*, 2003.

[58] V. Traag and J. Bruggeman. Community detection in networks with positive and negative links. *PRE*, 80(3):036115, 2009.

[59] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *KDD*, 2011.

[60] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *ICDM*, 2010.

[61] D. J. Watts. Computational social science: Exciting progress and future directions. *Winter Issue of The Bridge on Frontiers of Engineering*, 2014.

[62] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *CEAS*, 2008.

[63] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[64] R. Xiang and J. Neville. Collective inference for network data with copula latent markov networks. In *WSDM*, pages 647–656. ACM, 2013.

[65] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, 2010.

[66] X. Xu, N. Yuruk, Z. Feng, and T. A. Schweiger. Scan: a structural clustering algorithm for networks. In *KDD*, 2007.

[67] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, 2010.

[68] S.-H. Yang, A. J. Smola, B. Long, H. Zha, and Y. Chang. Friend or frenemy?: predicting signed ties in social networks. In *SIGIR*, 2012.

[69] M. Ye, X. Liu, and W.-C. Lee. Exploring social influence for recommendation: a generative model approach. In *SIGIR*, 2012.

[70] R. Zafarani, M. A. Abbasi, and H. Liu. *Social Media Mining: An Introduction*. Cambridge University Press, 2014.

[71] X. Zhang, J. Cheng, T. Yuan, B. Niu, and H. Lu. Toprec: domain-specific recommendation through community topic mining in social network. In *WWW*, 2013.

[72] S. Zhu, K. Yu, Y. Chi, and Y. Gong. Combining content and link for classification using matrix factorization. In *SIGIR*, 2007.

[73] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li, and Q. Yang. Discovering spammers in social networks. In *AAAI*, 2012.

[74] D. Watts, and S, Steven. Collective dynamics of 'small-world'networks. In *nature*, 1998.