Structured Learning from Heterogeneous Behavior for Social Identity Linkage

Siyuan Liu, Shuhui Wang, Feida Zhu

Abstract—Social identity linkage across different social media platforms is of critical importance to business intelligence by gaining from social data a deeper understanding and more accurate profiling of users. In this paper, we propose a solution framework, HYDRA, which consists of three key steps: (I) we model heterogeneous behavior by long-term topical distribution analysis and multi-resolution temporal behavior matching against high noise and information missing, and the behavior similarity are described by multi-dimensional similarity vector for each user pair; (II) we build structure consistency models to maximize the structure and behavior consistency on users' core social structure across different platforms, thus the task of identity linkage can be performed on groups of users, which is beyond the individual level linkage in previous study; and (III) we propose a normalized-margin-based linkage function formulation, and learn the linkage function by multi-objective optimization where both supervised pair-wise linkage function learning and structure consistency maximization are conducted towards a unified *Pareto* optimal solution. The model is able to deal with drastic information missing, and avoid the curse-of-dimensionality in handling high dimensional sparse representation. Extensive experiments on 10 million users across seven popular social networks platforms demonstrate that HYDRA correctly identifies real user linkage across different platforms from massive noisy user behavior data records, and outperforms existing state-of-the-art approaches by at least 20% under different settings, and 4 times better in most settings.

Index Terms—Social identity linkage, structured Learning, heterogeneous behavior, multi-resolution temporal information matching

1 Introduction

The ability of assuming multiple identities has long been a dream for many people. Yet it is not until the late advent of online social networks that this ambition of millions has been made possible in cyber virtual world. In fact, the recent proliferation of social network services of all kinds has revolutionized our social life by providing everyone with the ease and fun of sharing various information like never before (e.g., micro-blogs, images, videos, reviews, location check-ins). Meanwhile, probably the biggest and most intriguing question concerning all businesses is how to leverage this big social data for better business intelligence. In particular, people wonder how to gain thorough understanding of each individual user from the vast amount of online social data records. Unfortunately, information of a user from the current social scene is fragmented, inconsistent and disruptive. The key to unleashing the true power of social media is to link up all the data of the same user across different social platforms, offering the following benefits to user profiling.

Completeness. Single social networks service offers only a partial view of a user from a particular perspective. Cross-platform user linkage would enrich an otherwise-fragmented user profile to enable an all-around understanding of a user's interests and behavior patterns.

Consistency. For various reasons, information provided by users on a social platform could be false, conflicting, missing and deceptive. Cross-checking among multiple platforms helps improve the consistency of user information.

Continuity. While social platforms come and go, the underlying real persons remain, and simply migrate to newer ones. User identity linkage makes it possible to integrate useful user information from those platforms that has over time become less popular, or even abandoned.

Towards automatic user identity linkage of the same natural person across different social media platforms, we study to construct statistical learning method based on massive online user behavior data records. The research challenges can be addressed from the following aspects.

Unreliable Attributes. How users register their names online varies among different platforms. For example, a user tends to add family name after "Adele" in English communities, and users are likely to put a Chinese name or bizarre characters before or after "Adele" for eccentricity in Chinese communities. To make things worse, people do not use their true names, women would not tell their true ages, and males even pretend to be females. Statistical models (e.g. SVM [1], [2], [3]) or rule based models [4], [5] constructed with mere username [1], [2] and attribute analysis are far from being robust for accurate user linkage across online social communities.

Data Misalignment. User data on different social platforms could be misaligned in various ways that makes it hard to measure the behavior similarity among users.

- *Platform Difference.* User behavior may be divergent and platform dependent. For example, users might post their opinions about "life of youth" on Facebook and their political views on Twitter. Our study on 5 million users

- *Siyuan Liu is with the Heinz College at Carnegie Mellon University. siyuan@cmu.edu*
- *Shuhui Wang is with the Key Lab of Intellectual Information Processing, Institute of Computing Technology, CAS, Beijing, China. wangshuhui@ict.ac.cn*
- *Feida Zhu is with the School of Information Systems, Singapore Management University. fdzhu@smu.edu.sg*

from five most popular Chinese social platforms and 5 million users from two most popular English social platforms reveals a 25% to 85% difference in user generated content between different platforms. Moreover, the user behavior can be represented by various types of media, e.g., locations, blogs, tweets, videos and images, which we refer to as *heterogeneous behavior* in this paper. The platform-dependent and heterogeneous behavior would lead to extremely low-quality information matching.

- *Behavior Asynchrony.* Even semantically similar actions could often exhibit significant temporal variance. For example, a user would post selected pictures from a trip on Facebook in a certain time period. At a different time, the same or different pictures from the trip may be posted by the user again on Twitter.
- *Data Imbalance.* There has been a huge imbalance in terms of data volume between a user's primary social account and the rest, while statistical learning on such imbalanced data record has remained a long standing problem in machine learning community.

Missing Information. Due to privacy considerations, users may deliberately hide certain pieces of information online. Our study on real social media data indicates that at least 80% of users are missing at least two profile attributes out of the six most popular ones, and merely 5% of users have all attributes filled up. Drastic information missing leads to great difficulty for data distribution modeling on the behavior feature space in the learning process.

The above mentioned issues pose two main challenges for linkage function learning. First, reliable attribute and behavior feature modeling of online users should be constructed to measure the similarity among users from their heterogeneous and noisy online behavior records. Second, the difficulties brought by drastic information missing and insufficient linkage information require new learning strategy which is able to take advantage of structure information (i.e., the frequently interacted friends of each user) to improve the model generality. Existing work have applied heuristic processing in the profile information such as partial username overlapping and solved the problem by a set of binary classification models [1], [2]. However, these methods may work well only when information is veracious the ground-truth labels are available. Moreover, the heuristics they rely on are not always valid among platforms of different languages and cultures, resulting in low recall and significant bias.

In this paper, we propose HYDRA, a framework for cross-platform user identity linkage via heterogeneous behavior modeling. Compared with the long studied record linkage problem [6], [5], our technical breakthrough comes from taking advantage of two important features unique to social data: (I) *user behavior trajectory along temporal dimension*, and (II) *user's core social networks structure*, which is the part formed by those closet to the user, and is called "core structure" for short. The intuition is that (I) both empirical and social behavior studies (e.g., [7]) demonstrate that, over a sufficiently long period of time, a user's social behavior exhibits a surprisingly high level of consistency across different platforms; and (II) a user's core structures across different platforms share great similarity and offer a highly discriminative characterization of the user.

Based on (I), we model the behavior similarity among online users with multi-dimensional similarity vectors with the following information: a) the relative importance of the user attributes, which measures how likely two users refer to one person when one of their attributes is identical; b) the statistical divergence of topic distribution, describing the potential inclination of users over a long period; c) the overall matching degree of the behavior trajectories, capturing the identical actions between user accounts over a certain period of time. Based on (II), we develop a linkage function learning methodology by jointly optimizing the pair-wise identity linkage with ground-truth linkage information and seeking the social structure level behavior consistency among users without ground-truth linkage information. The key intuition is to propagate the linkage information along the linked users and their social structures. Consequently, the linkage function can be effectively learned even with partial ground truth linkage information.

In summary, the key contributions are as follows.

1. Heterogeneous Behavior Model. We design a new heterogeneous behavior model to measure the user behavior similarity from all aspects of a user's social data. It is able to robustly deal with missing information and misaligned behavior by long-term behavior distribution construction and a multi-resolution temporal behavior matching paradigm.

2. Structure Consistency. We propose a novel structure modeling method to maximize the behavior consistency on the users' core structure instead of user level behavior similarity. By propagating the linkage information along the social structure of each individual user, our model is capable of identifying user linkage even when ground-truth labeled linkage information is insufficient.

3. Multi-objective Model Learning. We solve the social identity linkage problem by multi-objective optimization (MOO) framework [8], where both the supervised learning on ground truth linkage information and the cross-platform structure consistency maximization are jointly performed towards a Pareto optimality. Specifically, we modify the formulations of kernel and linkage function, and develop a normalized-margin-based approach to deal with information missing in the similarity modeling. Theoretical analysis shows that our model is a generalized semi-supervised learning framework.

4. Experiments on Large-scale Real Data Sets. We evaluate HYDRA against the state-of-the-art on two real data sets — I) five popular Chinese social networks platforms and (II) two popular English social networks platforms — a total of 10 million users on 7 social media platforms amounting to more than 10 tera-bytes data. Experimental results demonstrate HYDRA outperforms existing algorithms in identifying true user linkage across different platforms.

2 Related work

User Linkage across Social Media. User linkage was firstly formalized as connecting corresponding identities across communities in [9] and a web-search-based approach was proposed to address it. Previous research can be categorized into three types: user-profile-based, user-generated-content-based, user-behavior-model-based and social-structure-based. User-profile-based methods collect tagging information provided by users [10], [11] or user profiles from several social networks and then represent user profiles in vectors, of which each dimension corresponds to a profile field [12], [13], [14]. Methods in this category suffer from huge effort of user tagging, different identifiable personal information types from site to site, and privacy of user profile. User-generated-content-based methods [1], on the other hand, collect personal identifiable information from public pages of user-generated content. Yet these methods still make the assumption of consistent usernames across social platforms, which is not the case in large-scale social networks platforms. User-behavior-model-based methods [2] analyze behavior patterns and build feature models from usernames, language and writing styles. Social-structure-based user linkage conduct linkage analysis by using structure features in social circles [15], [16], [17], [18]. For example, Korula et al. [15] solve the reconciliation of user's social network by starting from nodes with high degrees. Koutra et al. [17] formulates the user linkage problem by learning an optimal permutation function between two graph affinity matrices. Based on user's social, spatial, temporal and text information, Kong et al. [16] propose Multi-Network Anchoring to find the links between users from different platforms. Zhang et al. [18] propose to predict heterogeneous links (social links and location links) inside the target social network given a set of anchor links among users from target network and source network. Previous methods 1) seldom handle the missing information in usernames, user-generated content, behaviors and social structure; and 2) have not given interpretation why there exists such missing information and how it impacts the user linking result.

Authorship Identification across Documents. Authorship identification is a task that identifies the authors of documents by their writing and language styles analyzed from their corresponding documents. Previous studies on authorship identification can be categorized into two types: content-based and behavior-model-based. Content-based-methods identify content features across a large number of documents [19], [20], [21]. Behavior-model-based methods capture writing-style features [4], or build language models [22] to identify content authorship. However, different from document scenario, social media platforms are much more complicated with multiple data media, graph/ social structures and missing information, which compromises most authorship identification methods.

Entity Resolution across Records. User linkage is in one way or another related to problems from other research communities including co-reference resolution in natural language processing [23], entity matching [24], graph node classification [25], record linkage in database [6], [5], and name disambiguation in information retrieval [26], [27], which can be generalized as entity resolution across records. Different from previous structure-based feature extraction approach [25] and single feature based approaches [6], [5], we consider a much more challenging setting where we examine multiple features along time-line with missing and misaligned information and multiple media environments to link users across different platforms. Similarly, previous work on user identification on single site and de-anonymization in social networks have been surveyed in [1], [2], which are not elaborated here.

3 Problem Definition and Overview

Denote as P the set of all natural persons in real life. For a social networks platform S, denote as C_{S} the set of all usernames each belonging to a distinct user and $\phi_{\mathsf{S}} : C_{\mathsf{S}} \mapsto \mathsf{P}$ the injective function mapping each online user of S to a natural person.

Definition 1. ***Social Identity Linkage (SIL)****: Given two social networks platforms S and S', the problem of Social Identity Linkage (SIL) is to find a function f to decide if any two users from S and S' respectively correspond to the same natural person, i.e., $f : C_{\mathsf{S}} \times C_{\mathsf{S}'} \mapsto \{0, 1\}$ such that for any pair of users $(u_i, u_{i'}) \in C_{\mathsf{S}} \times C_{\mathsf{S}'}$, we have*

$$f(u_i, u_{i'}) = \begin{cases} 1 \quad , & if\ \phi_{\mathsf{S}}(u_i) = \phi_{\mathsf{S}'}(u_{i'}) \\ 0 \quad , & otherwise \end{cases} \tag{1}$$

It is worth noting that the straightforward approach to solve the problem by examining each pair of users would entail a high computational cost. Given an *SIL* problem instance of two social networks platforms S and S' with N_1 and N_2 users respectively, the number of all possible functions f by considering all the possible numbers of matched users is:

$$\sum_{n=1}^{\min(N_1,N_2)} \frac{N_1! N_2!}{n!(N_1-n)!n!(N_2-n)!} \tag{2}$$

where $N! = \prod_{k=1}^{N} k$. When we consider *SIL* problem on more platforms, the search space of pair-wise examination grows exponentially with the number of different platforms. Therefore, by only employing the user level pair-wise linkage information, huge amount of ground-truth linked pairs are required for training. From statistical linkage function learning aspect, it means that we need to collect statistically sufficient samples from the real data, to guarantee the convergence to the globally optimal linkage model. On the other hand, strong consistency is observed in behavior pattern and inclination among the users in the strongly interacted social friend groups [7], [28], [2], i.e., the stronger the interaction among users, the more similar their behavior and inclinations are. This observation endows us with the possibility of alleviating the difficulty by seeking the structure consistency of the candidate linked pairs generated by simple behavior matching across platforms. Furthermore, by joint optimization of the pair-wise linkage model and structure consistency, the linkage model can reach its full potential as the ground-truth linkage information can be reliably propagated along the user core social structure step by step, and finally a robust linkage model can be firmly constructed.

By taking the above mentioned issues into consideration, we propose HYDRA, a user linkage framework based on multi-objective optimization. It is composed of three main steps.

Step 1. Behavior Similarity Modeling. We calculate similarity among pairs of users via heterogeneous behavior modeling. Details are discussed in Section 4.

Step 2. Structure Information Modeling. We construct the structure consistency graph on user pairs by considering both the core network structure of the users and their behavior similarities. Details are discussed in Section 5.

Step 3. Multi-objective Optimization with Missing Information: We construct multi-objective optimization which jointly optimizes the prediction accuracy on the labeled user pairs and structure consistency measurements across different platforms. The model is further modified to deal with significant information missing. Details are discussed in Section 5.

We consider three kinds of data for model learning: (1) labeled data, including ground-truth linked pairs and pre-matched pairs, and the rest are (2) unlabeled pairs with no linkage information. The pre-matched labeled data is generated by our rule-based filtering, a much more sophisticated set of measures than existing methods, including partial username overlapping [1], [2], user attribute matching and user profile image matching by face recognition techniques[29]. by rule-based filtering. By combining heterogeneous behavior modeling and user core social networks structure, together with labeled data, into a multi-objective optimization, our approach conducts *SIL* on groups of users by taking full advantage of the context and content from social media.

4 Heterogeneous Behavior Model

The key challenges in modeling user behavior across different social media platforms are (I) *the heterogeneity of user social data* and (II) *the temporal misalignment of user behavior across platforms*. The high heterogeneity of user social data can be appreciated by the following categorization of all the data about a user available on a typical social platform.

1) **User Attributes.** Included here are all the traditional structured data about a user, e.g., demographic information, contact, etc. (Subsection 4.1).
2) **User Generated Content (UGC).** Included here are the unstructured data generated by users such as text (reviews, micro-blogs, etc.), images, videos and so on. Modeling is primarily targeted at *topic* (Subsection 4.2) and *style* (Subsection 4.3).
3) **User Behavior Trajectory.** User behavior trajectory refers to all the social behavior of a user as exhibited on the platforms along the time-line, e.g., befriend, follow/unfollow, retweet, thumb-up/thumb-down, etc. (Subsection 4.4).
4) **User Core Social Networks Features.** A user's core social networks are the social networks formed among those who are the closet to the user, and the features are the aggregation of the user's core social networks behavior (Subsection 4.5).

4.1 User Attribute Modeling

Textual Attributes. The profile information is informative in distinguishing different users. Common textual attributes in a user profile include name, gender, age, nationality, company, education, email account, etc. A simple matching strategy can be built on such a set of information. However, the relative importance of these attributes are not identical, because attributes such as gender and common names like "John" are not as discriminative as others such as email address in identifying user linkage. Yet, the weights of the attributes used in the matching can be learned from large training set by probabilistic modeling.

Specifically, given a set of N labeled training user pairs from different platforms, the relative importance of the attributes can be estimated by data counting. For a specific attribute a_k, $k = 1, ..., M_A$, we estimate the relative importance score by the following equation:

$$m_t(k) = \frac{P_D(k)}{P_D(k)+N_D(k)}, \;\; m_t(k) = \frac{m_t(k)+\varepsilon}{\sum\limits_{k'=1}^{M_A} m_t(k')+M_A\varepsilon} \tag{3}$$

where $P_D(k)$ represents the number of user pairs matched on a_k in the positive labeled set P_D, and $N_D(k)$ represents the number of pairs matched on a_k in the negative labeled set N_D. ε denotes a small real number that avoids over-fitting. M_A denotes the number of attributes. If a_k is missing for user i or i', it is denoted as a missing feature.

Given a user pair, an exact M_A dimensional attribute matching feature can be calculated. For example, if the user pair (i, i') is matched on 1st, 2nd, and 5-th attributes, where the corresponding weight of them are 0.1, 0.3, and 0.2, respectively, then the attribute feature of the user pair is $[0.1, 0.3, 0, 0, 0.2, ...]$. If any k-th attribute of user i or i' is absent, we denote the k-th feature as missing.

Visual Attributes. Besides textual attributes, visual attributes such as face images used in the profile can also be used to link users. However, as many users may not use their true face images, or use those with poor illumination and severe occlusion, such information could be very noisy. We designed a matching scheme to safely compare two user profile images. The work flow of face matching can be referred to [30]. In particular, if faces have been detected from both images, the pre-trained classifier is used to determine if the two faces correspond to the same person. As a standard work flow in face identification, we use the face detection approach, facial feature extraction and face classifier in [29]. For a large scale image processing, we implement the face attribute matching in a distributed computing environment, so that the matching of pairs of face images can be performed in parallel.

4.2 User Topic Modeling

An important feature of social media platform is that over a sufficiently long period of time, the UGC of a user collectively gives a faithful reflection of the user's topical interest. Faking one's interests all the time defeats the purpose of using a social networks service. Therefore, we propose to model a user's topical interest by a long-term user topic model. We first construct a latent topic model using Latent Dirichlet Allocation by using the collected textual messages after a textual preprocessing procedure, and the output of which is a probability distribution in the topic space. The number of latent topics is set to 300 in our study. We then calculate the multi-scale temporal topic distribution within a given temporal range for a user using the multi-scale temporal division [30]. The intuition comes from the fact that if two users refer to one person, their inclinations tend to be similar in the whole temporal range. Moreover, their inclinations in a shorter time

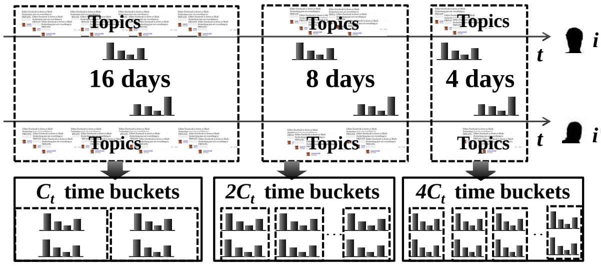


Fig. 1. Illustration of user topic modeling. First, the whole temporal range of user behavior data is divided into a set of time intervals with predefined values (e.g., 16 days, 8 days and 4 days). Then, all the distribution vectors within different time intervals are weighted and concatenated into one topic distribution vector. After that, the corresponding similarity of the topic distributions in each time interval and the whole range can be constructed. At last, the overall similarity between user i and i' is calculated as the similarities of all the time intervals, where a local matching is endowed with a larger weight than a global matching [30].

period should also be similar. The more their inclinations are locally matched in every shorter time period, the more similar their inclinations will be in the global range. Thus the users are more likely to be the same person.

Specifically, as shown in Figure 1, the time axis is divided into multiple time intervals with different scales (we use 1, 2, 4, 8, 16 and 32 days in this paper, which guarantees good performance). Based on the topic models, we obtain a topic distribution vector for each time interval, which is the average topic distribution for the user contents within this time interval. All the topic distribution vectors within each interval are accumulated into a single distribution, which represents the topic distribution pattern within this time interval. In Figure 1, C_t denotes the number of time intervals when the scale is selected to be 16. Correspondingly, the number of time intervals will be $2C_t$ and $4C_t$ respectively for 8 days and 4 days. Based on this, the similarity of temporal topic evolution of the specific scale between two users can be calculated by averaging the similarity of each temporal interval, where each similarity can be measured by the chi-square kernel or histogram intersection kernel [30]. Finally, all the similarities calculated using different time scales are concatenated into a weighted similarity vector, where local inclination matching will be endowed with a large importance than a global inclination matching.

The proposed long-term user topic models the behavior similarity on pair-wise topic correlation from coarse-to-fine resolutions. In this paper, we analyze the following distribution types using this proposed strategy:

Content Genre Distribution. The content genre measures the relevance between the textual messages and several popular topics on social media sites, e.g., sports/ music/ entertainment/ society/ history/ science/ art/ high-tech/ commercial/ politics/ geography/ traveling/ fashions/ digital game/ industry/ luxury/ violence, which are selected to cover the most popular topic genres.

Sentiment Pattern Distribution. According to studies on sentiment mining [31], [32], we can model the sentiment pattern using a two dimensional space (arousal-valence) [31] or roughly divide the emotion into several categories, e.g., happy/ fear /sad /neutral. It can be done by extracting the representative emotional key words in the textual content and learning a sentiment vocabulary. After that, each textual message can be represented by a probabilistic distribution on the sentiment vocabulary. We use the scheme in this section to construct the multi-scale similarity on sentiment pattern between two users.

4.3 User Style Modeling

The language style of a user including personalized wording and emotion adoption is usually well reflected in comments, tweets and re-tweets (e.g. function words extraction [1]), which is beneficial to distinguishing between different users. To model a user's characteristic style, we extract the most unique words of each user by a simple term frequency analysis on the whole database. Note that since the unique words may also be mistaken input, we can select the $k(k = 1, 3, 5)$ most unique ones after removing stop words from the least-used terms of the whole user data repository.

For user pairs, we can simply measure S_{lea}, their similarity on the unique word pattern, by word matching (the words should be uniformly converted, such as lower-case and singular form):

$$S_{lea} = \frac{\#matched\_words}{k} \tag{4}$$

4.4 Multi-resolution Behavior Modeling

User behavior trajectory is a unique feature of social media data laying out a user's behavior along the time line. In this paper, we are mainly concerned with the following patterns:

Location and Mobile Trajectory Information. Social media sites with location-based-service provide strong support and incentive for recording and sharing user locations. Generally, over an extended period of time, two users with mutually exclusive mobility patterns will not be the same person in reality. On the other hand, similar trajectory patterns across the platforms and no conflicting instances indicate the mobility similarity in real world, as they would like to provide check-in information on multiple social media platforms. By analyzing the mobility similarity over an long period, a sufficiently high similarity in mobile trajectory implies that the two users share similar and even exactly the same mobility behavior in real world. Therefore, the high mobility similarity can be considered as an important evidence in social identity linkage.

Multimedia Content Generation and Sharing. Users may post similar or duplicate multimedia content on the Web. For example, they may upload or share exactly the same image, video and music. However, if a high level of synchrony has been observed over an extended period of time between two users from different platforms, it is reasonable to hypothesize that these two users correspond to the same person.

A natural solution is to construct a set of pattern-matching sensors, one for each modality (location, visual, textual and audio), and use them to collectively evaluate user behavior similarity. However, as people are not always using multiple

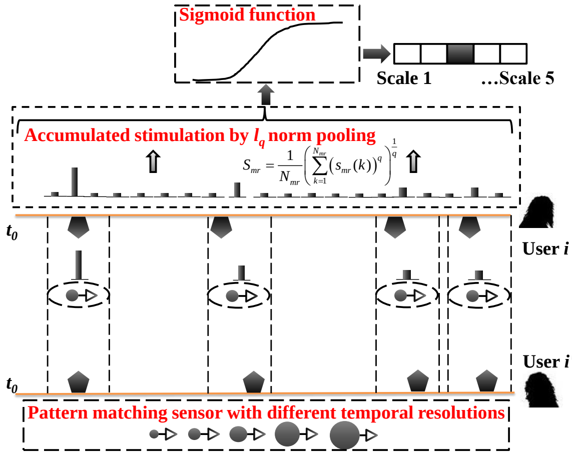

Sigmoid function
Scale 1
...Scale 5
Accumulated stimulation by l_q norm pooling
$S_{mr}=\frac{1}{N_{mr}}\left(\sum_{k=1}^{N_{mr}}(s_{mr}(k))^q\right)^{\frac{1}{q}}$
t_0
User i
User i'
t_0
Pattern matching sensor with different temporal resolutions

Fig. 2. Multi-resolution temporal behavior modeling. A set of pattern-matching sensors are designed. For two users, the sensors are used to detect the corresponding type of matched behavior within certain temporal scales. When all the matched behaviors have been detected by sensors, an l_q norm pooling and nonlinear sigmoid mapping aggregate all matched behavior signals into a multi-resolution similarity vector.

social platforms simultaneously, a significant amount of information could be missing in such a task. We therefore propose a multi-resolution temporal behavior model to perform pattern matching with the ubiquitous presence of missing information.

As shown in Figure 2, given two users i and i', we first construct a set of pattern-matching sensors with different temporal searching ranges. If there are patterns (denoted by pentagons) matched within the selected range of the pattern-matching sensor, it gives a positive stimuli signal. After we have collected all the stimuli signals along a certain period, we calculate the l_q-norm non-linear stimulation function, which is a trade-off between average pooling and max-pooling as:

$$S_{mr} = \frac{1}{N_{mr}}\left(\sum_{k=1}^{N_{mr}} (s_{mr}(k))^q\right)^{\frac{1}{q}}, q \geq 1 \qquad (5)$$

where $s_{mr}(k)$ denotes the score of k-th pattern matching sensor, S_{mr} represents aggregated behavior similarity, and N_{mr} represents the number of detected matched pattern. Next, we fit a sigmoid function to transform S_{mr} into a new stimulated signal $\widehat{S}_{mr} \in [0, 1]$. We repeat such processing with different pattern-matching sensors. Finally, a multi-dimensional pattern-matching feature is formed between user i and i', whose dimension is equivalent to the number of pattern-matching sensors.

Using l_q-norm is a natural choice from the bio-inspired stimulation. It has been found that the maximum stimulation from a pooled signal set will play significant role for perception. When q tends to be infinite, the signal selection tends to better approximate the maximum stimulation (i.e., max-pooling). Since the pattern-matching would be performed under different temporal scales, we can extract a multi-resolution temporal matching pattern between two users on the sparsely and asynchronously occurred patterns. The sigmoid function $\widehat{S}_{mr} = \frac{1}{1+e^{-\lambda S_{mr}}}$ is a typical nonlinear transformation function, where the parameter λ can be tuned on the specific validation dataset. Another important advantage of using l_q-norm is that it can reduce the number of dimensions in the behavior similarity construction by aggregating the sparsely matched patterns.

The pattern-matching sensors we construct in this paper are:

Location Matching Sensor. A location matching sensor calculates location adjacency by a Gaussian kernel on geo-coordinates of user i and user i' within the predefined spatial range [30].

Near Duplicate Multimedia Sensor. We construct a set of domain-specific duplicate content analysis models to detect the near duplicate multimedia content. We extract wavelet feature and cepstrum feature on each audio file, and then learn a support vector machine to decide if two audio files are duplicate. A spatial consistency graph model [33] is constructed for near duplicate image sensor. For near duplicate video detection, we apply [33] on each key video frame of video shot and develop a simple heuristic rule set for quick determination. Besides content analysis, the meta data (i.e., web address, time stamp and content providers) of each multimedia document can be used to quickly judge if they are duplicate.

The proposed framework can be further extended by designing and incorporating more special purpose detectors to capture the similarity from more content on online social platforms in future study.

4.5 Core Social Networks Features

It has been observed that, over time, users tend to bring their closest friends over to different social platforms they frequently use. Therefore, the behavior of a user's close friends are also informative in identifying different accounts of the same user. In [34], the average similarity of the neighborhood data of two data items is more robust compared with the original similarity since it calculates the similarity of two convex hulls instead of two data points. Inspired by [34], we model the behavior of a user's social connections. Given two users i and i' from different platforms, the behavior data of their top-k most frequently interacting friends are collected. For example, we denote their top-3 interacting friends as i_1, i_2, i_3, and i'_1, i'_2, i'_3, then the average behavior similarity and the standard deviation of the social connections of user i and i' can be calculated as:

$$\begin{aligned} S^1_{sc}(i,i') &= \frac{\sum\limits_{p=1}^{3}\sum\limits_{q=1}^{3} s(i_p, i'_q)}{9}, \\ S^2_{sc}(i,i') &= \sqrt{\frac{\sum\limits_{p=1}^{3}\sum\limits_{q=1}^{3}\left(s(i_p,i'_q) - S^1_{sc}(i,i')\right)^2}{9}} \end{aligned} \qquad (6)$$

where $s(i_p, j_q)$ denotes the similarity of any particular similarity measure described in previous sections. If we have 10-dim similarity description between user i and i', then a similarity vector with 30-dim is generated, including both the original similarity between i and i' (10-dim), the average neighborhood similarity (10-dim) and the standard deviation of their social connection (10-dim). The average similarity features and the standard deviation features measure the inclination and the behavior consistency of the friend groups, respectively.

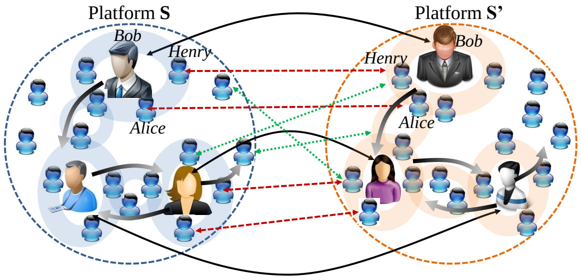


Fig. 3. Structure consistency maximization. Given two platforms, we measure both behavior similarity and structure consistency among their frequently communicating friends (elliptical rings), especially users with ground truth linkage information (linked by black arrows). The arrows within each platform indicate how the linkage information can be propagated along the social structure of each user. Consequently, the true linked user pairs (red dashed arrows) are correctly identified while the falsely linked user pairs (green dashed arrows) are filtered out.

5 Multi-objective Structure Learning

Based on the heterogeneous behavior modeling from user attributes, UGC and behavior trajectories as explained in Section 4, we propose to learn the linkage function via a multi-objective optimization framework.

Supervised Learning. Some social media platforms allow users to log in to different platforms with one account. For example, we can use a Facebook account to log in to Twitter. We collect such user-provided linkage information as the ground-truth label information. We notice that the labeled training pairs collected by our paradigm is much cleaner (precision over 95%) than the approach in [1] (precision around 75%) where the labeled training pairs are automatically generated based on the uniqueness (n-gram probability) of user names. We also collect label information by user attribute matching as the pre-linked label information. By utilizing the collected label information, we minimize the structured loss (SVM objective function) on the labeled training data.

Structure Consistency Modeling. We optimize the linkage function by maximizing both behavior similarity and social structure consistency between platforms. By constructing a positive semidefinite second-order structure consistency matrix among candidate linked user pairs, our model is able to consider the global structure between platforms to identify the true linkages and filter out those false ones, as illustrated in Figure 3. Most importantly, it compensates for the shortage of ground truth linkage information for user-level supervised learning by propagating the linkage information along the core social structure (i.e., friends with the most frequent interactions) of each individual user.

Multi-objective Optimization. We learn the linkage function by jointly minimizing the two objective functions via a unified multi-objective optimization framework. We prove that our model is a generalized semi-supervised learning approach by leveraging both ground truth linkage information and social structure.

5.1 Decision Model on Pairwise Similarity

Given a set $\mathbb{P}_l$ of N_l user pairs with ground-truth labels represented as: $\{(x_{ii'}, y_{ii'})\}$, where $x_{ii'}$ denotes the D-dimensional pair-wise similarity vector between user i and user i' calculated by the above behavior modeling methods, and $y_{ii'} \in \{1, -1\}$ denotes the label indicating whether the two users correspond to the same natural person. We denote the index set of user pairs with labels as $\mathbb{P}_l$. The decision model **f** to predict if a pair of users belong to the same natural person is represented as:

$$f(x) = \mathbf{w}^T x + b \tag{7}$$

where $\mathbf{w}$ and b are the model parameters that can be learned by minimizing the following objective function:

$$\begin{array}{l} F_D(\mathbf{w}) = \frac{\gamma_L}{2}||\mathbf{w}||^2 + \sum \xi_{ii'} \\ s.t.\ y_{ii'}(\mathbf{w}^T x_{ii'} + b) \geq 1 - \xi_{ii'} \end{array} \tag{8}$$

where $\xi_{ii'}$ denotes the slack variables that allow the model for non-linearly separable cases and b denotes the bias learned from the data. The optimization of objective function F_D is the standard structured risk minimization of binary classification.

5.2 Structure Consistency Modeling

The supervised learning relies heavily on a sufficient amount of ground truth linkage information. On the other hand, users' social structure information is an important complementary piece of information if its power in inferring user linkage is fully unleashed, as illustrated by the example in Figure 3. If $Alice$, Bob and $Henry$ are friends in real life, there would most likely be a high level of interaction frequency and behavior similarity among their corresponding accounts on the same platform. Such a consistent structure is indicated by the elliptic rings in Figure 3. A main strongly-connected cluster formed by correctly linked users (the dashed red arrows in Figure 3) would generate agreement links (edges with positive weights) among one another. These links are formed when the behavior of pairs of linked users agree at the level of social structure (their frequently interacting friends). Second, incorrect user linkage outside the cluster or weakly connected to it do not form strongly connected clusters due to the slim chance of establishing agreement links coincidentally (the dashed green arrows in Figure 3). When the ground truth linkage between the accounts of $Alice$ and $Henry$ is not available, we can still reliably link their accounts across the platforms based on the linked accounts of Bob together with the strong interaction observed from their social structures. Such linkage prediction can be further propagated to other frequently interacting friends of $Alice$ and $Henry$. Consequently, the linkage can be regularized towards the consistency at social structure level rather than individual user level.

To model the structure consistency, first, a set of candidate matched pairs are generated by measuring the behavior similarity between users i and i' from platform S and S', respectively, given two platforms S and S' containing N_{S} and $N_{\mathsf{S}'}$ users. For each candidate matching $a = (i, i')$, there is an associated affinity score that measures the similarity between user i and user i'. For each pair of assignments (a, b), where $a = (i, i')$ and $b = (j, j')$, there is an affinity score that measures how compatible the users (i, j) are with the users (i', j'). Given a list of candidate user pairs $\mathbb{P}_l \bigcup \mathbb{P}_u$, we store the affinities on every candidate $a \in \mathbb{P}_l \bigcup \mathbb{P}_u$ and every pair of candidate $a, b \in \mathbb{P}_l \bigcup \mathbb{P}_u$ in $\mathbf{M}$, such that (I) $M(a, a)$ is the affinity score measuring the individual-level similarity for candidate matching user pair $a = (i, i')$ based on the cross-platform behavior similarity. User pairs that are unlikely to be linked due to significant discrepancy in behavior patterns will be filtered out; (II) $M(a, b)$ is the affinity score measuring the similarity between user pairs $a = (i', j')$ and $b = (i, j)$ based on the pairwise behavior similarity as well as social structure consistency. $M(a, b) = 0$ if the inconsistency between (i, j) and (i', j') is too large. We assume $M(a, b) = M(b, a)$ without loss of generality.

We represent the agreement cluster C^* by an indicator vector y, such that $y(a) = 1$ if $a \in C^*$ and zero otherwise. The correspondent problem is reduced to find a cluster C^* of candidate user pairs (i, i') that maximizes the structure consistency $F_S(y) = \sum_{a,b \in C^*} M(a, b) = y^T \mathbf{M} y$. We relax both the mapping constraints and the integral constraints on y, such that its elements can take real values in $[0, 1]$. By the Raleigh's ratio theorem, the solution that maximizes the inter-cluster score $y^T \mathbf{M} y$ is the principal eigenvector of M.

By defining the relation between y and $\mathbf{w}$ as $y(ii') = \mathbf{w}^T x_{ii'}$, maximizing $F_S(y)$ is equivalent to the following problem:

$$\begin{array}{l} \min\limits_{\mathbf{w}} F_S(\mathbf{w}) {=} \mathbf{w}^T X^T \left(\mathbf{D} - \mathbf{M}\right) X \mathbf{w} \\ s.t. \ ||\mathbf{w}||^2 \leq s, D(a, a) = \sum_b M(a, b) \end{array} \qquad (9)$$

where s is a predefined real positive number which is used to prevent the norm of $\mathbf{w}$ from being arbitrarily large.

For users from C social platforms, we can decompose the problem into a set of one-to-one *SIL* problems with respect to $\mathbf{M}^{cc'}$, where $c \leq c', c = 1, ..., C-1$ and $c' = 2, ..., C$, without much effort. Then, the objective function $F_S(\mathbf{w})$ can be extended to an objective function vector $\mathbf{F}_S(\mathbf{w}) = [F_S^{cc'}(\mathbf{w})]$. The structure consistency matrix $\mathbf{M}^{cc'}$ is constructed as follows. First, for each candidate user pair $a = (i, i')$, their behavior similarity is calculated by $M^{cc'}(a, a) = \exp\left(\frac{-||x_i - x_{i'}||^2}{\sigma_1^2}\right)$, where σ_1 denotes the bandwidth to control the sensitivity on behavior similarity. Second, for candidate user pair $a = (i, i')$ and $b = (j, j')$, their structure consistency is calculated by:

$$\begin{array}{l} M^{cc'}(a, b) = \exp\left(\frac{-\left(||x_i - x_{i'}||^2 + ||x_j - x_{j'}||^2\right)}{2\sigma_1^2}\right) \cdot \\ \left(1 - \frac{(d_{ij} - d_{i'j'})^2}{\sigma_2^2}\right) \end{array} \qquad (10)$$

where σ_2 denotes the bandwidth to control the structure sensitivity of user social relations. d_{ij} denotes the n-hop distance measuring the closeness of two users, which is formally defined as the minimal number of friends (including the user himself) that a user reaches the friend user. Specifically, we define k_{ij} as the number of intermediate users from user i to j, and then their distance is $d_{ij} = (k_{ij} + 1)^2$.

It is not hard to prove that matrix $\mathbf{M}^{cc'}$ is positive-definite, and consequently, matrix $\Theta^{cc'} = \mathbf{D}^{cc'} - \mathbf{M}^{cc'}$ is positive-semidefinite by spectral graph theory. Details are omitted due to space limit.

Our model is consistent with [35] that the majority of user's friends tend to provide useful information besides the users themselves. Other works follow similar rules with diversified assumptions on the data structure. For example, in [15], the graph reconciliation is started on a set of nodes with high degrees. In "Big-Align" model [17], the optimal permutation is learned among the adjacency matrices of two graphs. Our model is similar in spirit of the Big-Align model because both use structural matching for linking social identities. However, the adjacency information in our model is constructed by joint modeling of behavior and social circles, instead of only considering the follower/followee information.

5.3 Multi-objective Optimization

Based on the two above-mentioned objective functions (F_S and F_D), given C social platforms and their users, we formulate the *SIL* problem as a multi-objective optimization problem [8]:

$$\begin{array}{l} \min\limits_{\mathbf{w}} F(\mathbf{w}) = [F_D(\mathbf{w}), \mathbf{F}_S(\mathbf{w})] \\ s.t. \ c_{ii'}(\mathbf{w}^T x_{ii'} + b) \geq 1 - \xi_{ii'}, i \in \mathsf{S}, i' \in \mathsf{S}', ||\mathbf{w}||^2 \leq s \end{array} \qquad (11)$$

where $F(\mathbf{w})$ denotes a $(C-1)C/2+1$ dimensional objective function vector.

A feasible solution does not typically exist that minimizes all objective functions simultaneously in such a problem. Note that since a penalty on the squared norm of $\mathbf{w}$ has been included in F_D, then constraint $||\mathbf{w}||^2 \leq s$ can be omitted. Therefore, we define a *Utility function* to aggregate all the objective functions in form of generalized weighted exponential sum as:

$$U = \sum_{k=1}^{(C-1)C/2+1} w_k [F_k(\mathbf{w})]^p, \forall k, F_k(\mathbf{w}) > 0, w_k \geq 0 \qquad (12)$$

where the weight parameter w_k is a preference parameter encoding decision makers' preference. By minimizing *Utility function* U, we seek the Pareto optimal solutions [8], which cannot be improved in any of the objectives without degrading at least one of the other objectives.

Proposition 1. *The solution of the weighted exponential sum utility function U is sufficient and neccessary for Pareto optimality.*

Proof: See Athan *et. al.* [36] and Yu [37]. □

When $p = 1$, the utility function is similar with traditional semi-supervised learning objective function with a weighted combination of empirical loss, the penalty on $\mathbf{w}$ and a graph Laplacian regularizer [38]. When $p > 1$, our model is viewed as minimizing the distance function between the solution point and *Utopia* points [36] in the multi-dimensional objective function space.

Dual Problem. By introducing a nonlinear mapping $\phi(\cdot)$ to a higher (possibly infinite) dimensional Hilbert space $\mathbb{H}$. $\mathbf{w}$ and b define a linear regression in that space. According to the Representer Theorem [39], the decision function $\mathbf{w}$ can be expressed in the dual problem as the expansion over labeled user pairs and unlabeled candidate user pairs $\mathbf{w} = \sum\limits_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \alpha_{ii'} \phi(x_{ii'})$. Then, the decision function is given by:

$$f(x_t) = \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \alpha_{ii'} K(x_{ii'}, x_t) + b \tag{13}$$

where we use $\mathbf{K}$ to denote the kernel matrix formed by kernel functions $K(x_{ii'}, x_{jj'}) = \langle \phi(x_{ii'}), \phi(x_{jj'}) \rangle$. Take $p = 1$ as the illustrative example, by setting $w(1) = 1$ and $w(k) = \gamma_M, k = 2, ..., (C-1)C/2 + 1$, we plug Eqn. 13 into Eqn. 12 and introduce the *Lagrangian* multipliers, and obtain the following regularized *Utility function* to be minimized:

$$\begin{aligned} \min_{\alpha,\beta} \Big\{ & \tfrac{1}{2} \alpha^T \left(2\gamma_L \mathbf{K} + \tfrac{2\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} \mathbf{K}(\mathbf{D} - \mathbf{M})\mathbf{K} \right) \alpha \\ & -\alpha^T \mathbf{KJY} \beta + \beta^T \mathbf{1} \Big\} \end{aligned} \tag{14}$$

where β denotes an N_l-dimensional *Lagrangian* parameter vector, $\mathbf{J} = [\mathbf{I}, \mathbf{0}]$ is an $N_l \times |\mathbb{P}_l \bigcup \mathbb{P}_u|$ with $\mathbf{I}$ as the $N_l \times N_l$ identity matrix (the first N_l pairs are labeled) and $\mathbf{Y} = diag\{y_1, ..., y_{N_l}\}$. $\mathbf{M}$ denotes the cross-platform structure consistency matrix:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}^{12} & 0 & ... & 0 \\ 0 & ... & ... & ... \\ ... & ... & \mathbf{M}^{cc'} & 0 \\ 0 & ... & 0 & ... \end{bmatrix} \tag{15}$$

where $c < c', c = 1, ..., C-1, c' = 2, ..., C$. Similarly, $\mathbf{D}$ is the diagonal matrix. We obtain the solution by taking derivatives w.r.t. α:

$$\alpha = \left(2\gamma_L \mathbf{I} + 2\tfrac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} (\mathbf{D} - \mathbf{M}) \mathbf{K} \right)^{-1} \mathbf{J}^T \mathbf{Y} \beta^* \tag{16}$$

Again, substituting Eqn. 16 into the dual functional Eqn. 14, we obtain the following smooth quadratic programming problem to be solved:

$$\begin{aligned} & \beta^* = \max_{\beta} \left\{ \beta^T \mathbf{1} - \tfrac{1}{2} \beta^T \mathbf{Q} \beta \right\} \\ & s.t. \sum_{ii' \in \mathbb{P}_l} \beta_{ii'} y_{ii'} = 0, 0 \le \beta_{ii'} \le \tfrac{1}{|\mathbb{P}_l|} \end{aligned} \tag{17}$$

where:

$$\mathbf{Q} = \mathbf{YJK} \left(2\gamma_L \mathbf{I} + 2\frac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} (\mathbf{D} - \mathbf{M}) \right)^{-1} \mathbf{J}^T \mathbf{Y} \tag{18}$$

From the above derivation we can see that the *SIL* problem can be well cast into a standard convex programming problem that can be easily solved by many off-the-shelf optimization package. Despite that we only introduce the model construction procedure when $p = 1$, similar derivation can also be performed when $p > 1$. Consequently, the resulted objective function is also convex due to the convexity of the individual objective functions and the convexity of the *Utility function* U. Besides, using higher values for p increases the effectiveness of the method in providing the complete Pareto optimal set [36], [8].

Dealing with Information Missing. When constructing the pair-wise similarity among users, significant information missing exists among real world Web platforms due to: (A) intrinsically heterogeneous information sources, (B) unpredictable user login and logout, and (C) privacy concerns. Previous approaches [1], [2] construct discriminate models where the missing feature is automatically filled with zeros based on the assumption that the values do exist but not observed, which is actually not the case of the problem in hand. In this paper, we suppose that the missed values can not be observed, and they need not be filled with any value. We revise the discriminative model $f(x)$ by a normalized margin as [40]:

$$f(x) = \mathbf{w}^{ii'} \phi(x_{ii'}) + b, \mathbf{w} = \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \frac{\alpha_{ii'}}{s_{ii'}} \phi(x_{ii'}) \tag{19}$$

where $\mathbf{w}^{ii'}$ denotes the instance specific vector obtained by taking the entries of $\mathbf{w}$ that are relevant to $x_{ii'}$ $(\phi(x_{ii'}))$, namely, those for which the sample $x_{ii'}$ $(\phi(x_{ii'}))$ has valid features. $s_{ii'} = \sqrt{\frac{||\mathbf{w}^{ii'}||^2}{||\mathbf{w}||^2}}$ is a normalized scalar that can be estimated iteratively, where:

$$\begin{aligned} ||\mathbf{w}||^2 &= \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \sum_{jj' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \tfrac{\alpha_{ii'} \alpha_{jj'}}{s_{ii'} s_{jj'}} \langle \phi(x_{ii'}), \phi(x_{jj'}) \rangle \\ ||\mathbf{w}^{ii'}||^2 &= \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \sum_{jj' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \tfrac{\alpha_{ii'} \alpha_{jj'}}{s_{ii'} s_{jj'}} \langle \phi(x_{ii'}), \phi(x_{jj'}) \rangle_{R_{ii'}} \end{aligned} \tag{20}$$

where $\langle \cdot, \cdot \rangle_{R_{ii'}}$ denotes the kernel calculation using only the non-missing indices of user pair ii'. Consequently, the objective dual problem in Eqn. 17 is revised as:

$$\mathbf{Q} = \mathbf{YJK}_n \left(2\gamma_L \mathbf{I} + 2\frac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} (\mathbf{D} - \mathbf{M}) \right)^{-1} \mathbf{J}^T \mathbf{Y} \tag{21}$$

where each element of $\mathbf{K}_n$ is calculated by polynomial kernel:

$$K_n(x_{ii'}, x_{jj'}) = \frac{\left(\langle x_{ii'}, x_{jj'} \rangle_R + 1 \right)^d}{s_{ii'} s_{jj'}} \tag{22}$$

with the inner product calculated over valid features $\langle x_{ii'}, x_{jj'} \rangle_R = \sum\limits_{r: r \in R_{ii'} \cap R_{jj'}} x_{ii'}(r) x_{jj'}(r)$. It is straightforward to see that $\mathbf{K}_n$ is a kernel, since user pairs with missing values can be filled with zeros. In summary, the details of the model are described in Algorithm 1. The proposed model is guaranteed to converge to a local optimal solution within 5 iterations, according to experiments in this paper and in [40].

Since the data size would be extremely large, we adopt the distributed convex optimization method [41] to optimize the objective function distributively on several servers in parallel with a carefully designed model synchronization strategy. In summary, the sketch of the optimization process is described in Algorithm 1.

Algorithm 1 The HYDRA algorithm

```
Input: Data: X,Y, Parameters: $\gamma_L$, $\gamma_M$, $p$, $\sigma_S$, $\sigma_D$
Output: $\alpha$, $\beta$
Select the candidate pair set $\mathbb{P}_u$ by comparing the pair-wise similarity.
Construct structure consistency graph $\mathbf{M}$.
Initialize $s$ for all the labeled and unlabeled training pairs as: $s^0_{ii'} = 1 - \#absent\_feat/\#feat\_dim$.
while the stopping criterion is not reached do
    Find $\beta^t$ by Eqn. 17, and calculate $\alpha^t$ by Eqn. 16.
    Update $s^t$ based on $\mathbf{w}(t)$, using Eqn. 20, $t = t + 1$.
end while
```

5.4 Model Analysis

Interaction of multiple objectives. We learn the linkage function via optimizing two kinds of objective functions, i.e., the supervised learning using the reliably obtained ground truth, and the structure consistency maximization by modeling the core social networks behavior consistency. They are complementary to each other by jointly measuring the behavior similarity of both individual and group levels. When the ground truth information is insufficient (e.g., less than 10% of the pairs assigned with labels), the model will be more dependent on the core social networks structure. The linked user pairs will be served as some "*anchor*" pairs where the linkage information can be propagated along the core social networks. However, the learned model tends to be over-smooth (under-fitting) by over-emphasizing the structure consistency. When the ground truth information is sufficient (e.g., more than 80% of the training pairs assigned with labels), the model can still be endowed with more generalization power by the decision boundary smoothing towards better group level behavior consistency. The l_p-norm in the *Utility function* determines the way how the two kinds of objective functions interact with each other, where large p imposes more uniqueness on the dominant objective function. Correspondingly, model overfitting is likely to take place. Therefore, a better trade-off can be steadily achieved by appropriately tuning ω_k and p on different behavior data record repositories from different communities.

Complexity. We briefly analyze the model complexity of HYDRA. In Eqn. 17, our model achieves an $O(|\mathbb{P}_l|^2)$ time complexity, where $\mathbb{P}_l$ denotes the number of user pairs with ground truth linkage information. In Eqn. 18 and Eqn. 21, the time complexity is $O(s^2|\mathbb{P}_l \bigcup \mathbb{P}_u|^3)$, where s indicates the sparse level of the matrix $\mathbf{M}$. In fact, $\mathbf{M}$ is extremely sparse under real situations. For example, in our study, the sparse level of $\mathbf{M}$ is only about $[0.0001, 0.001]$ on Chinese social platforms. Therefore, the actual time consumption for Eqn. 18 and Eqn. 21 will be far less than the linkage function construction in Eqn. 17.

6 EXPERIMENTAL EVALUATION

Real Data. We use two publicly available large-scale real data sets for our experiments. The first one, referred to as "*Chinese*", includes five popular social networks services which were originated from China and have since gained global popularity. (1) **Sina Weibo:** (www.weibo.com) A hybrid of Twitter and Facebook with a user base of 500 million users and 47 million daily active users by December 2012. (2) **Tecent Weibo:** (t.qq.com) Another twitter-like micro-blogging service with 500 million users and over 100 million daily active users. (3) **Renren:** (www.renren.com) A social networks service dubbed as the Facebook of China with 162 million registered users. (4) **Douban:** (www.douban.com) A social networks service for people to share content on topics of movies, books, music, and other off-line events in Chinese cities, with over 100 million monthly unique visitors. (5) **Kaixin:** (www.kaixin001.com) A social networks service with 160 million registered users. We use 5 million users in this data set, each with accounts on every one of the five platforms. The time span of this data set is from June 2012 to June 2013. The second one, referred to as "*English*", includes two globally popular social networks: (1) **Twitter** (twitter.com); and (2) **Facebook** (www.facebook.com). We use 5 million users in this data set each with accounts on both Twitter and Facebook. The time span of this data set is from June 2012 to June 2013. For the above social networks, we collect user profiles (e.g. gender, city, and favorites), social content (e.g. tweets, posts, and status), social connections (e.g., friendship, comments, and repost or retweet contents), timeline information (e.g. time index for each behavior). Our ground truth of the linkage of each user across all the platforms are provided by a third-party data provider who has access to each Chinese user's national ID number, IP address and home address used by the user to register all accounts on different websites, all of which collectively serve as the most reliable data to uniquely identify a natural person and link all the different accounts. Note that users in the English data set are all Chinese users of our choice. In the following experiment results, x-axis is all decreasing ranked result (user is by degree, and community is by size). The ratio between the labeled data to unlabeled data is set to $1/5$, but we have also tested other ratio settings in our experiment.

Experiment Environment. Our experiments and latency observations are conducted on 5 standard servers (Linux), with Intel (R) Xeon (R) Processor E7-4870 (30M Cache, 2.40 GHz, 6.40 GT/s Intel (R) QPI, 10 cores), 64 GB main memory and 10,000RPM server-level hard disks.

Compared Methods. We compare both our methods with the following state-of-the-art approaches and our own baselines.

(I) MOBIUS: a behavior-modeling approach to link users across social media platforms [2].

(II) Alias-Disamb: an unsupervised data-driven approach based on username analysis to link users across platforms [1].

(III) SMaSh: a record linkage approach finding linkage points over Web data [5].

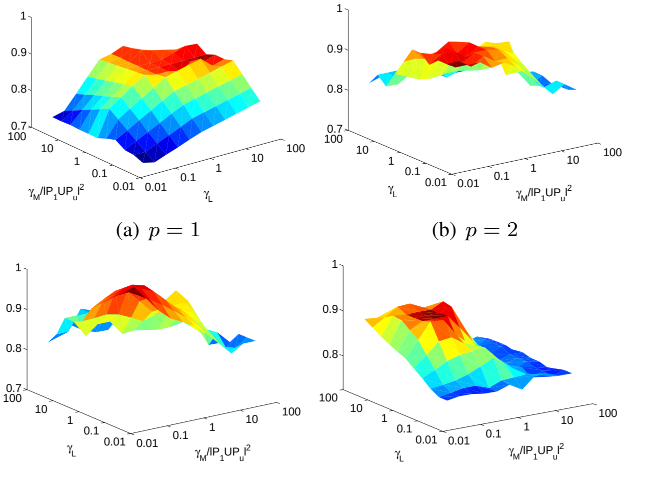
(IV) SVM-B: binary prediction on user pairs using support vector machines on the proposed similarity calculation schemes.

(V) HYDRA-Z: a degenerate version of our model HYDRA where all the missing features are filled with zeros (Eqn. 17).

(VI) HYDRA-M: our model HYDRA with missing features properly handled (Eqn. 21 and Algorithm 1). Without specification, we call HYDRA-M as HYDRA.

Parameter Settings. To achieve better performance of all the approaches, a validation set with 5 million user pairs and their ground truth labels have been used. For other compared learning-based linkage function learning, the model parameters are set to be the optimal value through the validation set. All of them are implemented on the distributed computing platforms as our approach. Specifically, for SVM-B, we also use a distributed optimization to learn the linkage model.

For pair-wise similarity calculation in this paper, parameters (e.g., ε for user profiling, q and λ for multi-resolution temporal similarity modeling) are tuned by a grid search procedure to maximize the performance of a linear SVM on validation set. The optimized multi-dimensional similarity $x_{ii'}$ are used for model construction of (IV), (V) and (VI).



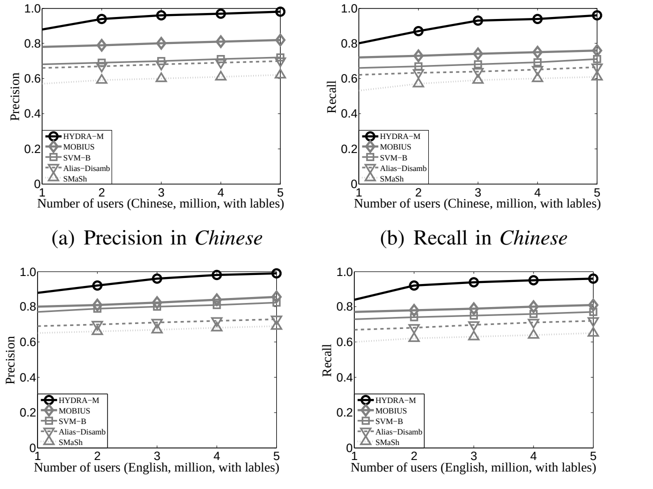
(a) $p = 1$

(b) $p = 2$

(c) $p = 3$

(d) $p = 4$

Fig. 4. Performance curve with different settings of γ_M and γ_T under different p.



(a) Precision in *Chinese*

(b) Recall in *Chinese*

(c) Precision in *English*

(d) Recall in *English*

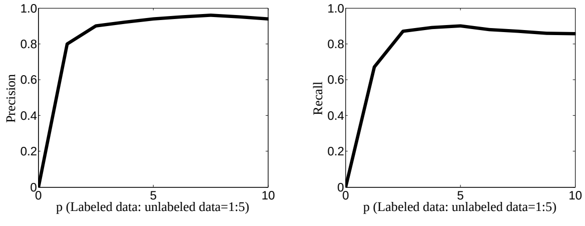
Fig. 5. Performance w.r.t. #labeled pairs.

For both HYDRA-Z and HYDRA-M, we need to tune the model parameters γ_L, γ_M, p, σ_S and σ_D. We construct the models on the training data and conduct parameter tuning on the validation set. In the following sections, we will illustrate the functional properties with respect to different model parameter settings.

Evaluation Metrics. In our experiments, we use precision and recall to evaluate the effectiveness, and the total execution time (at different scales) to evaluate the efficiency. Precision is defined as the fraction of the user pairs in the returned result that are correctly linked. Recall is defined as the fraction of the actual linked user pairs that are contained in the returned result. Parameters of all the kernels for HYDRA are tuned according to the methods described in the previous sections.

6.1 Effectiveness Evaluation

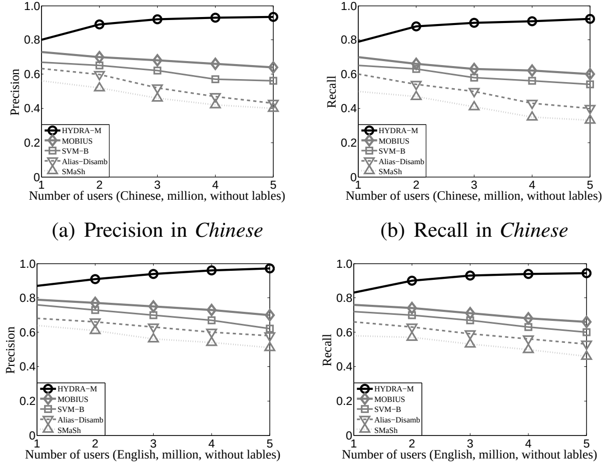
Performance w.r.t. Different γ_M and γ_L. We compare the performance of our approaches with different settings of γ_M and γ_L under $p = 1, 2, 3, 4$, and show the performance curves in Figure 4. From Section 5 we see that γ_M and γ_L determine the relative importance of the problems in MOO framework from the decision maker's perspective, while p determines how the learned model approximates the *Utopia* solution, thus determining the intrinsic structure of the *Utility* function.



(a) Precision

(b) Recall

Fig. 6. The precision and recall curve w.r.t. different p.



(a) Precision in *Chinese*

(b) Recall in *Chinese*

(c) Precision in *English*

(d) Recall in *English*

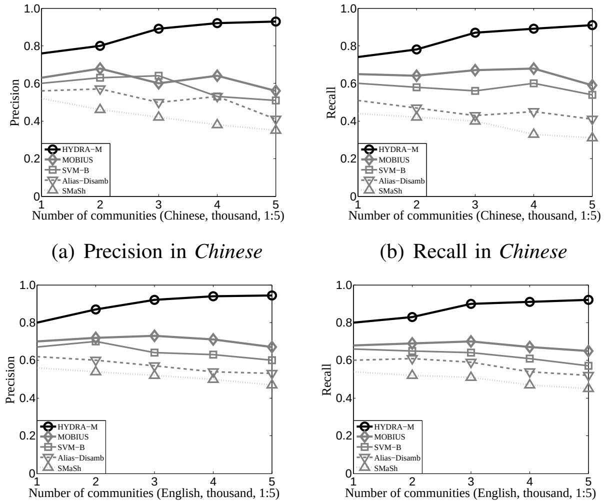
Fig. 7. Performance w.r.t. #unlabeled pairs.

However, for real data, a decision maker's preference does not necessarily correspond to the best performance, as can be seen from Figure 4. The results tell us that different settings of p lead to the choice of different optimal setting of γ_M and γ_L.

The performances in Figure 4 under different p indicate that a reasonable setting of γ_L is in $[0.01, 1]$. For γ_M, the optimal setting of the normalized value $\frac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2}$ may depends on the average number of friends for each user on different social platforms, where a reasonable setting should be in $[0.1, 10]$.

Performance w.r.t. Different p. Figure 6 shows our performance with p varied from $p = 1$ to $p = 10$ and the optimal setting of γ_M and γ_L. Although increasing p will help obtain the complete Pareto optimal solution, it does not necessarily correspond to the optimal solution of our *SIL* problem. In fact, imposing larger p leads to heavier preference on objective functions with larger values, leading inevitably to model overfitting. We see from Figure 6 that both precision and recall reach optimum with an appropriate setting of p ($p = 6$ and $p = 5$ for best precision and recall, respectively).

Performance w.r.t. Different Number of Labeled Pairs. Fixing the level of structure information, we vary the number of labeled user pairs from one million to five million users. The experimental results are reported in Figure 5. Note that, although the performance of all five methods shows improvement along with the increasing number of labeled pairs, the improvement of HYDRA's is the most significant and exhibits noticeably greater acceleration compared to the baseline methods. Another interesting observation is that the performance on English platforms are better than Chinese ones, which is also true for Figure 7. Our interpretation of it goes as follows. First, the complexity of the *SIL* problem grows with the number of platforms involved — we used five Chinese platforms and only two for English platforms. Second,



(a) Precision in *Chinese*

(b) Recall in *Chinese*

(c) Precision in *English*

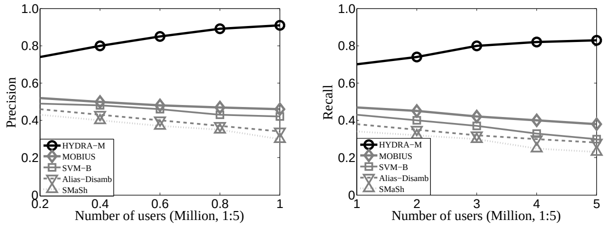
(d) Recall in *English*

Fig. 8. Performance with #social communities.

the social structure and behavior on Chinese platforms are characterized with a higher complexity and greater temporal dynamics than those on English platforms. We use real data to illustrate this in Figure 10 (comparing Twitter and Sina Weibo as an example). In Figure 10 (a), we plot the diffusion speed for retweets. In comparison, Sina Weibo has much more retweets and a higher diffusion speed for retweet than Twitter, which means the information diffusion in Sina Weibo is much faster than in Twitter. Combined with Figure 10 (b), the retweet distribution, we can tell that Sina Weibo contains much richer and more dynamic information than Twitter, presenting a much more challenging task for the *SIL* problem. In Figure 10 (c) and Figure 10 (d) [1], we plot the follower and followee distribution. Note that most users in Sina Weibo have much more followers and followees than in Twitter. Consequently, the much more complicated social structure contributes to the greater challenge to the *SIL* problem on Chinese platforms.

Performance w.r.t. Different Structure Information Levels. Fixing the number of labeled user pairs, we vary the numbers of user pairs with no ground truth labels, and evaluate the linkage precision. The results are illustrated in Figure 7. Compared against Figure 5, we notice that the performance of baseline methods with unlabeled data is much worse than the performance with labeled data in Figure 5. But our HYDRA survives the unlabeled data setup and performs much better than the baseline methods. In Figure 5 and Figure 7, HYDRA not only performs much better (higher precision and recall) than the baseline methods, but also shows better performance along the increasing number of users.

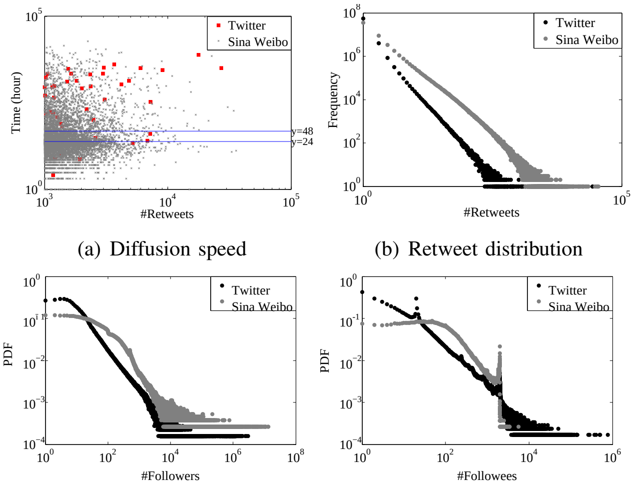
Performance w.r.t. Different Number of Social Communities. We evaluate how the structure information from other social communities [28] could help enhance the model generalization power. Specifically, given the top five largest overlapping communities A, B, C, D, E with labeled training pairs between A and B. To judge whether a user pair from $C_A \times C_B$ corresponds to the same person, we incrementally incorporate structure information of training pairs from $C_A \times C_C$, $C_A \times C_D$, $C_A \times C_E$, $C_B \times C_C$, $C_B \times C_D$ and $C_B \times C_E$ for model training, and report the results on the test set of user pairs from $C_A \times C_B$ in Figure 8. The interesting observation is that the social community structure has much greater impact on the results for Chinese platforms than those for English. It may due to the more complicated social community structure and social behaviors, as we have illustrated in Figure 10. But as we notice in Figure 8, the social community structure indeed helps HYDRA achieve better results than baseline methods.



(a) Precision

(b) Recall

Fig. 9. Performance on different social platforms.



(a) Diffusion speed

(b) Retweet distribution

(c) Follower distribution

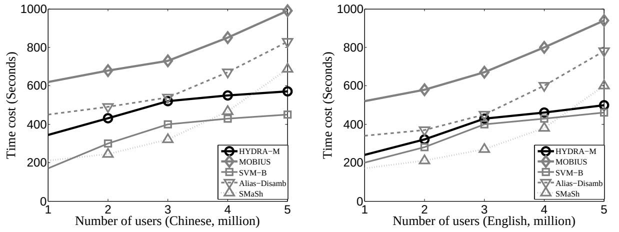
(d) Followee distribution

Fig. 10. Comparison between Twitter and Sina Weibo.

Performance w.r.t. Different Social Platforms. We study *SIL* across culturally different social platforms, that is, between Chinese platforms and English platforms. In this experiment, we use the whole data set with all seven different social networks. The results are reported in Figure 9. Compared with the previous results, there is an obvious performance drop (affected by different writing styles in Chinese and English, and social friends), but HYDRA performs even better than the baseline methods, and has better performance improvement with the increasing number of users. This shows that heterogeneous behavior model demonstrates better fitting to online social behaviors and social structure modeling helps to capture more linkable information.

Based on effectiveness evaluation in different parameter settings, different number of labeled data pairs, different structure information levels and different number of social communities, we conclude that HYDRA significantly outperforms the baseline methods and displays good scalability with the increasing amount of data, in no matter Chinese platforms or English platforms, or together.

[1] Spikes in Figure 10 (d) are due to the default setup of Twitter and Sina Weibo whereby (1) every Twitter new user are recommended with 10 followees by default (the left spike); and (2) For Twitter and Sina Weibo, there are a 2,000-followee limit (the right spike).



(a) *Chinese* (b) *English*

Fig. 11. Efficiency evaluation.

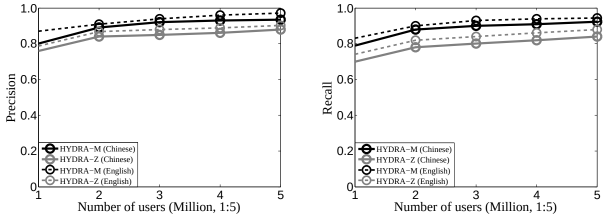
6.2 Efficiency Evaluation

We use the total execution time at different scales to evaluate the efficiency. From the results reported in Figure 11, HYDRA consumes less time than the baseline methods (except SVM-B and SMaSh) in the same scaling-up number of users, for both Chinese and English platforms. Since HYDRA solves a convex optimization problem where a unique global optimal solution can be achieved. It is interesting that the runtime cost of HYDRA increases at a slower speed than the baseline methods. Along with the scaling-up number of users, the runtime of HYDRA displays a converging tendency, which is a desirable feature for handling large-scale data sets. The explanation for this favorable characteristics lies in the the social structure we incorporate into the HYDRA model — for such five-million-user social networks, when we have accumulated around three million users and their one-hop friends, the social structure is almost well-constructed, and after that, the resulting utility function contains a rather sparse structure consistency matrix $\mathbf{M}$ which is easy to solve with many accelerating techniques (e.g., accelerated coordinate descent method). For Alias-Disamb [1], it automatically generates a large number of training pairs by analyzing the uniqueness of the usernames, where most of the generated label information may be incorrect, resulting in an extremely large quadratic programming problem and extremely slow convergence rate. SVM-B corresponds to one of the objective functions in our MOO learning framework, and it therefore consumes less time for model construction. SMaSh employs a totally different paradigm for record linkage. As a result, the property of its efficiency behavior is quite different from other discriminative-model-based approaches for *SIL*.

Our model possesses $O(|\mathcal{P}_l|^2)$ time complexity in learning the linkage function. However, when increasing the number of users to million scales, the portion of inactive users will be dominated. The behaviors of these users are very random and sparse so that the behavior similarities are almost zeroes among these users. For a linkage model learning, these inactive users will not become the "support vector", thus they will be reduced in the first several iteration rounds. Therefore, increasing the number of inactive users will not significantly increase the training time consumption. In conclusion, HYDRA is capable of handling large-scale data sets.

6.3 Sensitivity Evaluation

Sensitivity evaluation is to test HYDRA-M and HYDRA-Z under varied missing information settings (from varied number of users). According to the results in Figure 12, for both Chinese and English platforms, HYDRA-M outperforms HYDRA-Z although both achieve high precision and recall. The results clearly demonstrate the superiority of HYDRA-M (HYDRA) in handling missing information without compromising performance.



(a) Precision (b) Recall

Fig. 12. Performance with missing data.

6.4 Discussion

The data sizes in this paper are prohibitively large for a single PC or server. Despite that we deal with trillions of data records millions of users when optimizing the convex problem in Eqn. 17, the problem still can be handled efficiently by several servers by the following reasons.

First, the meaningful behavior patterns are extremely sparse. For example, the percentage of the non-missing or non-zero features on the data of English communities are no more than 4%, and the percentage of available similarities between users are no more than 2%. Even some missing values can be filled by aggregating the core social behavior similarities, the available similarities are still no more than 3%. Similarly, the available similarities are about 2% on the data of Chinese communities. The structure consistency matrices $\mathbf{M}$ is even more sparse, which is usually less than 1% non-zero elements for both English and Chinese communities. Such data sparsity allows efficient data storage, and successfully execution of our learning algorithm with 5 high-end servers.

Moreover, we learn the model by the distributed optimization method [41] which optimizes the linkage function in parallel on several servers with a carefully designed synchronization strategy. The core idea of the distributed optimization is that the overall objective function can be optimized towards the optimal solution via optimizing a series of sub-problems on different part of data stored distributively on different servers. Meanwhile, our model involves support vector representation, i.e., α and β, where at least 90% of the dimensions in β are zeros on million scale data. For each step of the model optimizing, we perform a coefficient space shrinking process to actively identifying the non-zero dimensions in β with a very simple gradient thresholding technique. Consequently, the corresponding entries with zeros in β in all the matrices (e.g., $\mathbf{M}$ and $\mathbf{K}$) can be excluded from the memory when optimizing β^t. Finally, we further enhances the efficiency of the model learning by using β^t as the warm start for optimizing β^{t+1}.

7 Conclusion

In this paper, we link user accounts across different social networks platforms. To deal with the challenges, we propose a framework, HYDRA, a multi-objective learning framework incorporating heterogeneous behavior model and core social networks structure. We evaluate HYDRA against the state-of-the-art on two real data sets. Experimental results demonstrate that HYDRA outperforms existing algorithms in identifying true user linkage across different platforms.

Acknowledgments. This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA) and the Pinnacle Lab at Singapore Management University, and was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, and in part by National Natural Science Foundation of China: 61303160. The authors also thank Jinbo Zhang and Ramayya Krishnan for valuable discussions and support.

REFERENCES

[1] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: an unsupervised approach to link users across communities," in *WSDM'13*, 2013.

[2] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in *KDD'13*, 2013.

[3] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migration patterns in social media," in *AAAI'11*, 2011, pp. –1–1.

[4] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the Association for Information Science and Technology*, vol. 57, no. 3, 2006.

[5] O. Hassanzadeh, K. Q. Pu, S. H. Yeganeh, R. J. Miller, M. Hernandez, L. Popa, and H. Ho, "Discovering linkage points over web data," *PVLDB*, vol. 6, no. 6, pp. 444–456, 2013.

[6] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data*, pp. –1–1, 2007.

[7] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, "Time-critical social mobilization," *Science*, vol. 334, no. 6055, pp. 509–512, 2011.

[8] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and multidisciplinary optimization*, vol. 26, no. 6, pp. 369–395, 2004.

[9] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *ICWSM'09*, 2009, pp. –1–1.

[10] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *ICWSM'11*, 2011, pp. –1–1.

[11] P. Jain and P. Kumaraguru, "@i to @me: An anatomy of username changing behavior on twitter," *CoRR*, 2014.

[12] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *ASONAM'12*, 2012, pp. 1065–1070.

[13] A. Nunes, P. Calado, and B. Martins, "Resolving user identities over social networks through supervised learning and rich similarity features," in *SAC'12*, 2012, pp. 728–729.

[14] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks," in *NDT'09*, 2009, pp. 360–365.

[15] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *PVLDB*, pp. 377–388, 2014.

[16] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *CIKM'13*, 2013, pp. 179–188.

[17] D. Koutra, H. Tong, and D. Lubensky, "Big-align: Fast bipartite graph alignment," in *ICDM'13*, 2013, pp. 389–398.

[18] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *WSDM'14*, 2014, pp. 303–312.

[19] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Record*, vol. 30, no. 4, pp. 55–64, 2001.

[20] E. Amitay, S. Yogev, and E. Yom-Tov, "Serial sharers: Detecting split identities of web authors," in *PAN'07*, 2007, pp. –1–1.

[21] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by compression," *IEEE Transactions on Information Theory*, pp. 1523–1545, 2005.

[22] J. Novak, P. Raghavan, and A. Tomkins, "Anti-aliasing on the web," in *WWW'04*, 2004, pp. 30–39.

[23] J. Cai and M. Strube, "End-to-end coreference resolution via hypergraph partitioning," in *COLING'10*, 2010, pp. 143–151.

[24] J. Wang, G. Li, J. X. Yu, and J. Feng, "Entity matching: How similar is similar," *PVLDB*, pp. 622–633, 2011.

[25] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos, "It's who you know: graph mining using recursive structural features," in *SIGKDD'11*. ACM, 2011, pp. 663–671.

[26] Y. nan Qian, Y. Hu, J. Cui, Q. Zheng, and Z. Nie, "Combining machine learning and human judgment in author disambiguation," in *CIKM'11*, 2011, pp. 1241–1246.

[27] D. V. Kalashnikov, Z. Chen, S. Mehrotra, and R. Nuray-Turan, "Web people search via connection analysis," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1550–1565, 2008.

[28] W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang, "Online search of overlapping communities," in *SIGMOD Conference'13*, 2013, pp. 277–288.

[29] http://www.briancbecker.com/bcbcms/site/proj/facerec/fbextract.html.

[30] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling." SIGMOD, 2014.

[31] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, pp. 143–154, 2005.

[32] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, pp. 55–64, 2003.

[33] L. Chu, S. Jiang, S. Wang, Y. Zhang, and Q. Huang, "Robust spatial consistency graph model for partial duplicate image retrieval," *IEEE Transactions on Multimedia*, 2013.

[34] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. Noble, "Semi-supervised protein classification using cluster kernels," *Bioinformatics*, pp. 55–64, 2005.

[35] F. Kooti, N. O. Hodas, and K. Lerman, "Network weirdness: Exploring the origins of network paradoxes," *arXiv:1403.7242v1*, 2014.

[36] T. W. Athan and P. Y. Papalambros, "A note on weighted criteria methods for compromise solutions in multi-objective optimization," *Engineering Optimization*, vol. 27, pp. 155–176, 1996.

[37] P.-L. Yu, Y.-R. Lee, and A. Stam, *Multiple-criteria decision making: concepts, techniques, and extensions*. Plenum Press New York, 1985.

[38] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[39] B. Schlkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond," *Cambridge: TheMITPress*, 2002.

[40] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller, "Max-margin classification of data with absent features," *Journal of Machine Learning Research*, pp. 1–21, 2008.

[41] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

Siyuan Liu is Research Scientist at Carnegie Mellon University. He received his first Ph.D. degree from Department of Computer Science and Engineering at Hong Kong University of Science and Technology, and the second Ph.D. degree from University of Chinese Academy of Sciences. His research interests include mobile data analytics and heterogeneous social networks mining.

Shuhui Wang received B.S. degree in Electronic Engineering from Tsinghua University, China, and Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, China. He is Assistant Professor with Institute of Computing Technology, Chinese Academy of Sciences and with Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include large-scale Web data mining, and visual semantic analysis.

Feida Zhu is Assistant Professor at School of Information Systems of Singapore Management University. He obtained Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign and B.Sc. in Computer Science from Fudan University, China. His research interests include large-scale data mining, graph/network mining and social network analysis.