HYDRA: Large-scale Social Identity Linkage via Heterogeneous Behavior Modeling

Siyuan Liu *, Shuhui Wang +, Feida Zhu #, Jinbo Zhang $, Ramayya Krishnan *
*Heinz College, Carnegie Mellon University
+Key Lab of Intellectual Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
#School of Information Systems, Singapore Management University
$School of Electronics Engineering and Computer Science, Peking University

ABSTRACT

We study the problem of large-scale social identity linkage across different social media platforms, which is of critical importance to business intelligence by gaining from social data a deeper understanding and more accurate profiling of users. This paper proposes HYDRA, a solution framework which consists of three key steps: (I) modeling heterogeneous behavior by long-term behavior distribution analysis and multi-resolution temporal information matching; (II) constructing structural consistency graph to measure the high-order structure consistency on users' core social structures across different platforms; and (III) learning the mapping function by multi-objective optimization composed of both the supervised learning on pair-wise ID linkage information and the cross-platform structure consistency maximization. Extensive experiments on 10 million users across seven popular social network platforms demonstrate that HYDRA correctly identifies real user linkage across different platforms, and outperforms existing state-of-the-art algorithms by at least 20% under different settings, and 4 times better in most settings.

Categories and Subject Descriptors: H.2.8 Database applications: Data mining

General Terms: Algorithms; Experimentation.

Keywords: User linkage, multiple information, social networks, heterogeneous behavior model.

1. INTRODUCTION

The recent blossom 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., microblogs, 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 a deeper and better understanding of each individual user from the vast amount of social data out there. 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 analysis is to link up all the data of the same user across different social platforms, offering the following benefits for user profiling.

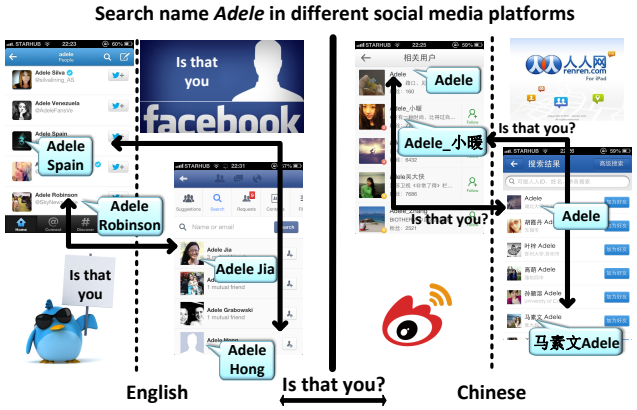


Figure 1: Ambiguity between users on different social communities. The left figure illustrates possible linked users between Twitter and Facebook. The right figure illustrates possible linked users between two popular Chinese platforms, Sina Weibo and Renren.

Completeness. Constrained by the features and design of each, any single social network 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.

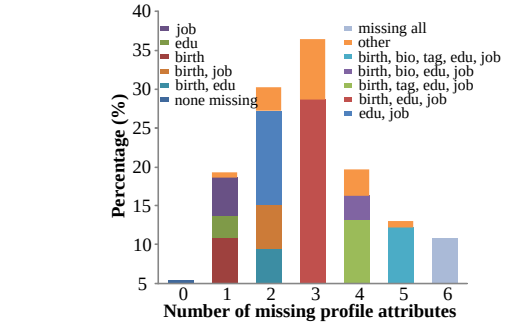
Consistency. For various reasons, information provided by users on a social platform could be false, conflicting, missing and deceptive. Cross-checking among multiple platforms helps improve the consistency of user information.

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

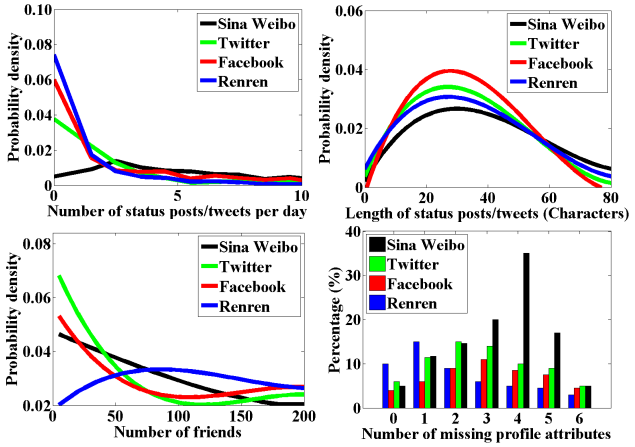
In this paper, we study the problem of automatically linking user accounts belonging to the same natural person across different social media platforms. It is beneficial to first explore the research challenges for a better understanding of this problem.

1.1 Research Challenges

Unreliable Usernames. How users register their names online varies among different platforms. Taking Figure 1 as an example, while a user tends to add family name after "Adele" in English



(a) Missing information



(b) Platform difference

Figure 2: Challenges of cross-platform social identity linkage. (a) Statistics of missing information in social media data (seven different platforms). (b) The online heterogeneous behavior represented by multiple media, where the similarity between users cannot be identified due to the platform difference.

communities, the user could be very likely to put a Chinese name before or after "Adele" in a Chinese community. To make things worse, some users may even add bizarre characters for eccentricity on certain platforms. Traditional approaches that heavily rely on username parsing [16, 32] to link users may fail on more diversified communities. On the other hand, statistical models (e.g. SVM [16, 32, 15]) or rule based models [33, 11] constructed with mere username and attribute analysis are far from being robust to accurately identify user linkage across online social communities.

Missing Information. Due to privacy considerations, users may deliberately hide certain pieces of information online. Figure 2 (a) shows our study on real social media data. 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.

Misaligned Data. One main challenge is that user data on different social platforms could be misaligned in various ways.

- *Information Veracity.* Users may deliberately provide false data on selected platforms. For example, many people do not use their true names, some women would not tell their true ages, and some males even pretend to be females.
- *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 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 on multiple platforms would lead to extremely low-quality information matching. (Figure 2 (b)).
- *Behavior Asynchrony.* Even semantically similar actions often exhibit significant temporal variance. For example, a user posts 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 again on Twitter.
- *Data Imbalance.* There has been observed a huge imbalance in terms of data volume between a user's primary social account and the rest.

1.2 Our Contribution

To address these challenges, we propose HYDRA, a framework for cross-platform user identity linkage via heterogeneous behavior modeling. Compared with the record linkage problem long studied in database community [8, 11, 24], our technical breakthrough comes from taking advantage of two important features unique to social data: (I) *user behavior trajectory along temporal dimension*: both empirical and social behavior studies (e.g., [23]) demonstrate that, over a sufficiently long period of time, a user's social behavior exhibits a surprisingly high level of consistency across different platforms. (II) *user's core social network structure*: the part formed by those closet to the user, and is called "core structure" for short. 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 user accounts belong to the same person when any one of their attributes is identical; b) the statistical divergence of topic distributions, describing the potential inclination of users over a long period of time; c) the overall degree of matching between users' behavior trajectories, capturing the highly correlated actions between user accounts over a certain period of time. Based on (II), we develop a linkage function learning methodology by taking full advantage of the agreement of the social structure level behavior consistency. The key intuition is to propagate the linkage information based on the linked users and the strong interaction along their social structures. Consequently, the linkage function can be effectively learned even with partial ground truth linkage information. In summary, our key contributions are:

1. Heterogeneous Behavior Modeling. We design a new heterogeneous behavior model to measure the user behavior similarity from all aspects of a user's social data. The proposed framework 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 Modeling. We propose a novel social structure modeling method to leverage users' core social network structure to identify user linkage. We measure the high order pairwise behavior similarity and structure consistency by a graph representation. The model is learned to maximize the structure consistency, which is equivalent to a convex objective function minimization. By incorporating structure consistency, our model is capable of identifying user linkage even when ground-truth labeled linkage information is insufficient.

3. Multi-objective Model Learning. We put forward a multi-objective optimization (MOO) framework [19] to solve the overall social identity linkage problem, which jointly optimizes the supervised learning on labeled user linkage pairs and the cross-platform structure consistency maximization. To deal with missing information, we further enhance the model into an iterative learning one. We also provide theoretical proof that our model is a generalization of the traditional semi-supervised learning, and can be efficiently solved by convex optimization.

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

Road Map. Section 2 discusses related work. Section 3 formally defines the social identity linkage problem. Section 4 presents an overview of our approach. Section 5 presents our heterogeneous behavior model and Section 6 proposes the multi-objective model learning. Section 7 presents the experimental results on different data sets. Finally, Section 8 concludes the paper.

2. RELATED WORK

2.1 User Linkage across Social Media

User linkage was firstly formalized as connecting corresponding identities across communities in [31] 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 and user-behavior-model-based. User-profile-based methods collect tagging information provided by users [13] or user profiles from several social networks and then represent user profiles in vectors, of which each dimension corresponds to a profile field (e.g., username,profile picture,description,location,occupation, etc.) [18, 22, 27]. 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 [16], 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 network platforms. User-behavior-model-based methods [32] analyze behavior patterns and build feature models from usernames, language and writing styles. Unfortunately, previous methods 1) have not handled the missing information prevalent among usernames, user-generated content, user behavior and social structures; 2) have not explored the underlying reasons for the missing information and its impact on user identity linkage; 3) have not well formalized the user linkage problem with a solution of a sound theoretical foundation. To the best of our knowledge, our work is the first to link users across different social media platforms by integrating all the social data associated with a user in a unified model.

2.2 Authorship Identification across Documents

Authorship identification is a task that identifies the authors by analyzing their writing and language styles from their corresponding documents. Previous studies on authorship identification can be categorized into two kinds: content-based and behavior-model-based. Content-based-methods identify content features across a large number of documents [7, 5]. Behavior-model-based methods capture writing-style features [33], or build language models [21] to identify content authorship. However, different from the document setting, social media platforms are characterized by data of much greater heterogeneity, complicated network structures and a high degree of missing information, which could easily compromise most authorship identification methods.

2.3 Entity Resolution across Records

User linkage is also in one way or another related to problems from other research communities including co-reference resolution in natural language processing [4], inter-media data retrieval [26], entity matching [28], record linkage in database [8, 11, 24], and name disambiguation in information retrieval [20, 14], which can be generalized as entity resolution across different records. In contrast to previous studies, we consider the user linkage problem in a much more challenging setting where we examine multiple features along time-line with missing and misaligned information across multiple media platforms.

Also related are previous studies on user identification on a single site and de-anonymization in social networks, which have been well surveyed in [16, 32].

3. PROBLEM DEFINITION

Denote as P the set of all natural persons in real life. For a social network 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. Our social identity linkage problem is defined as follows.

DEFINITION 1. ***Social Identity Linkage (SIL)****: Given two social network platforms S and S', the problem of Social Identity Linkage (SIL) is to find a function f to decide if any two users picked 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 this problem by examining every user pair without any attribute filtering entails high computational cost. Given an *SIL* problem instance of two social network platforms S and S' with N_1 and N_2 users respectively, the number of all possible functions f we are to examine is given by:

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

where $N! = \prod_{k=1}^{N} k$.

When we try to solve the *SIL* problem on even more platforms, the search space size grows exponentially with the number of different platforms. A natural solution is to apply certain filtering to reduce the number of candidate user pairs. Existing work have applied heuristic knowledge on the profile information such as partial username overlap and solved the problem by a set of binary classification models [16, 32]. However, these methods may work well only when the ground-truth labels are available for training. Moreover, the heuristics they rely on are not always valid among platforms of different languages and cultures, resulting in low recall and significant bias.

We classify all user pairs in our data into three kinds: (1) ground-truth linked pairs; (2) pre-matched pairs by rule-based filtering; and

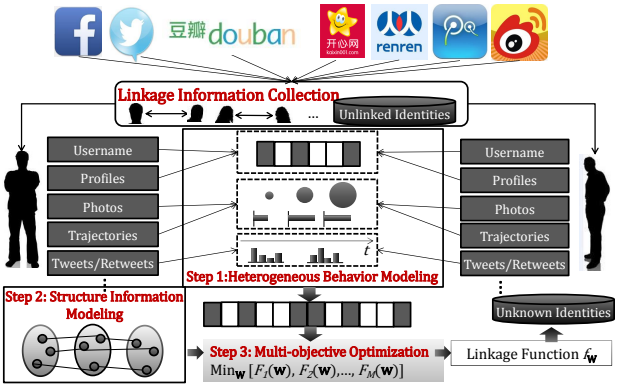


Figure 3: HYDRA framework.

the rest are called (3) unlabeled pairs. We call (1) and (2) together as labeled data. Our solution differs from previous work in the following two aspects.

1. We consider both labeled and unlabeled data in computing the linkage function. While the labeled training data plays a role, it does not work alone. The power of our solution lies in combining heterogeneous behavior modeling and user core social network structure, together with labeled data, into a multi-objective optimization, such that linkage can be identified even between two users of unlabeled pairs which are not pre-matched by rule-based filtering.
2. The pre-matched labeled data is generated by our rule-based filtering, which includes a much more sophisticated set of measures than existing methods, including partial username overlapping [16, 32], user attribute matching and user profile image matching by face recognition techniques[12].

4. FRAMEWORK OVERVIEW

In this paper, we propose HYDRA, which integrates both users' heterogeneous behavior and their core social network structure into a unified multi-objective user linkage framework. HYDRA is composed of the following three main steps as illustrated in Figure 3.

Step 1. Behavior Similarity Modeling. We calculate the similarity between two users of a pair for all user pairs via heterogeneous behavior modeling. Details are discussed in Section 5.

Step 2. Structure Consistency 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. Based on the previous two steps, we convert the *SIL* problem into a two-class classification problem and construct multi-objective optimization which jointly optimizes the prediction accuracy on the labeled user pairs and multiple structure consistency measurements across different platforms. Details are discussed in Section 6.

5. 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 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 information, etc. (Subsection 5.1)
2. **User Generated Content (UGC).** Included here are the unstructured data generated by users such as text (reviews, microblogs, etc.), images, videos and so on. Modeling is primarily targeted at *topic* (Subsection 5.2) and *style* (Subsection 5.3).
3. **User Behavior Trajectory.** User behavior trajectory refers to all the social behavior of a user exhibited on the platforms along the time-line, e.g., befriend, follow/unfollow, retweet, thumb-up/thumb-down, etc. (Subsection 5.4)

To address these two challenges, we propose a behavior modeling framework which computes the similarity between two users by capturing the heterogeneity in their behavior as well as the characteristics of their temporal evolution.

5.1 User Attribute Modeling

Textual Attributes. Common textual attributes in a user profile include name, gender, age, nationality, profession, education, email account, etc. While user profile information is effective in distinguishing different users, the relative importance of these attributes could be different, since attributes such as gender and popular names like "John" are not as discriminative in identifying user linkage as some others such as email address. Yet, the weights of the attributes used in the matching can be learned from large training data 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 of a_k by the following equation:

$$m_t(k) = \frac{P_D(k)}{P_D(k)+N_D(k)},\ m_t(k) = \frac{m_t(k)+\varepsilon}{\sum_{k'=1}^{M_A} m_t(k')+M_A\varepsilon} \quad (3)$$

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 used to avoid over-fitting. If a_k is absent for user i or i', it is denoted as a missing feature.

Visual Attributes. Besides textual attributes, visual attributes such as face images used in the profile can also help link users. However, such information could be very noisy as the face images might not be real, or come with poor illumination and severe occlusion. We design a matching scheme as shown in Figure 4 to compare two user profile images. In particular, if faces are detected from both images, the pre-trained classifier is used to determine if the two faces correspond to the same person. We use the face detector, facial feature extraction and face classifier provided by [12].

5.2 User Topic Modeling

An important feature of social media platform is that in general, over a sufficiently long period of time, the UGC of a user collectively gives a faithful reflection of the user's topical interests. Faking one's interests all the time defeats the purpose of using a social network service. Therefore, we propose to model a user's topical interests by a long-term user topic model. We first construct a latent topic model using Latent Dirichlet Allocation on every textual message, the output of which is a probability distribution over the topic space. We then calculate the multi-scale temporal topic distribution within a given temporal range for a user using the multi-scale temporal division similar to [17]. As shown in Figure 5, first, the time

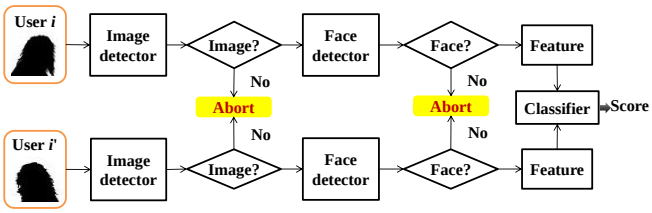


Figure 4: The workflow of face recognition for identity linkage. A face detector is employed to extract the face from a pair of profile images. Then a pre-trained face classifier outputs a confidence score in $[0, 1]$ indicating how likely the two faces belong to one person.

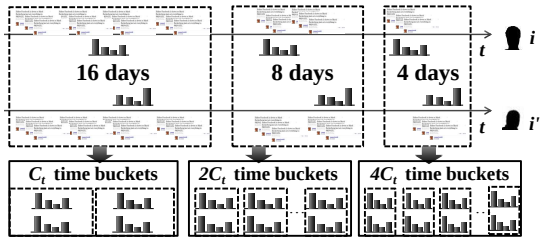


Figure 5: Illustration of user topic modeling. First, the temporal axis is divided into a series of time buckets with predefined scales (e.g., 16 days or 8 days). Then all the distribution vectors within a time bucket are aggregated into one topic distribution. After that, the corresponding similarity between the topic distributions in each time bucket can be constructed. Finally, the overall similarity between user i and i' is calculated by averaging over the similarities of all the time buckets.

axis is divided into multiple time buckets with different scales (we use 1, 2, 4, 8, 16 and 32 days in this paper to guarantee the optimal performance), then all the topic distribution vectors within each bucket are aggregated into a single distribution, which represents the topic distribution pattern within this time bucket. In Figure 5, C_t denotes the number of time buckets when the scale is selected to be 16. Correspondingly, the number of time buckets will be $2C_t$ and $4C_t$ respectively for 8 days and 4 days. Based on this, the similarity of topic evolution of a specific scale between two users can be simply calculated by averaging over the similarities of all temporal intervals, where each similarity can be measured by the chi-square kernel or histogram intersection kernel [17]. Finally, all the similarities calculated using different time scales are concatenated into a similarity vector.

The proposed long-term user topic model captures the behavior similarity from pair-wise topic correlation at a series of 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 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.

Sentiment Pattern Distribution. According to studies in sentiment mining [10], we can model sentiment patterns using a two-dimensional space (arousal-valence) [10] or roughly group all emotions into several categories, e.g., happy/ fear /sad /neutral. It can be done by extracting 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 compute a multi-scale similarity measure on sentiment patterns between two users.

5.3 User Style Modeling

The language style of a user including personalized wording and emoticon adoption is usually well reflected in comments, tweets and re-tweets (e.g. function words extraction [16]), 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 these unique words may also be inaccurate, we 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. Such choices of k have been adopted widely in related studies [29].

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

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

5.4 Multi-resolution Behavior Modeling

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

Mobile Trajectory and Location Information. Social media sites with location-based-service provide strong support and incentive for users to record and share their locations. Generally, users with similar trajectory patterns and no conflicting instances over an extended period of time are likely to be the same person in real life.

Multimedia Content Generation and Sharing. Users may post similar multimedia content on the web. For example, they may upload or share exactly the same image/ video/music. However, If a high level of synchrony is observed over an extended period of time between two user accounts 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 similarity. However, as people are not always using multiple 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 6, given two users i and i', we first construct a set of pattern-matching sensors with different temporal searching ranges. If matched patterns (denoted by pentagons) are identified within the selected range of a pattern-matching sensor, a positive stimuli signal would be generated. After we have collected all the stimuli signals along a certain time period, we calculate the l_q-norm non-linear stimulation function as follows:

$$S_{mr} = \frac{1}{N}\left(\sum_{i=1}^{N}(s_{mr}(i))^q\right)^{\frac{1}{q}}, q \geq 1 \tag{5}$$

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', with the number of dimensions the same as the number of pattern-matching sensors.

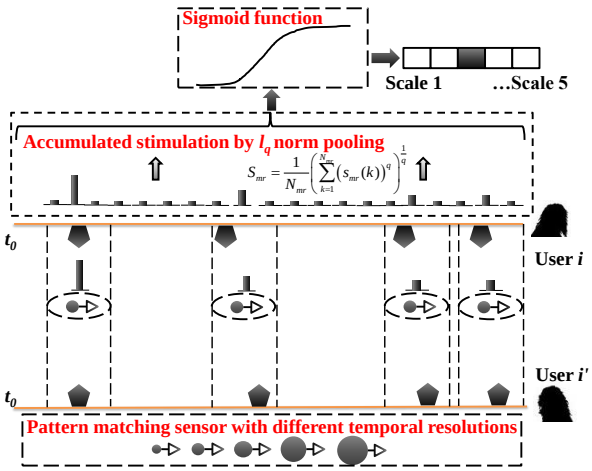


Figure 6: Multi-resolution temporal behavior modeling. A set of pattern-matching sensors are designed. For two users, sensors are used to detect the corresponding type of matched behavior within certain temporal scales. When all the matched behaviors have been detected by sensors, an l_q norm pooling and nonlinear sigmoid mapping then aggregates all matched behavior signals into a multi-resolution similarity vector.

The choice of l_q-norm is inspired by bio-stimulation. It has been found that the maximum stimulation from a pooled signal set plays a significant role for perception. When q approaches infinity, 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 occurring 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. The pattern-matching sensors we construct in this paper are the following:

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 [17].

Near Duplicate Multimedia Sensor. A near duplicated image sensor or down-sampling method [9] is constructed for near duplicate multimedia sensor.

6. MULTI-OBJECTIVE MODEL LEARNING

Based on the heterogeneous behavior modeling from user attributes, UGC and behavior trajectories as explained in Section 5, 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 [16] (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.

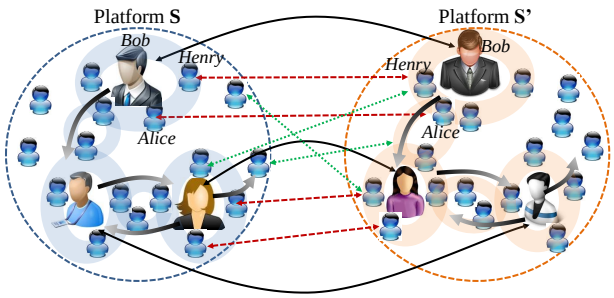


Figure 7: Structure consistency maximization. Given two platforms, we measure both behavior similarity and structure consistency among their most frequently communicating friends (elliptical rings), especially those with ground truth linkage information (linked by black arrows). The arrows within each platform indicate how the linkage information can be propagated along the social structure of each user. Consequently, the true user linkage (red dashed arrows) are correctly identified while the falsely linked user pairs (green dashed arrows) are filtered out.

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 7. 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.

6.1 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 $\mathbf{f}$ to predict if a pair of users belong to the same natural person is represented as:

$$f(x) = \mathbf{w}^T x + b \tag{6}$$

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

$$\begin{aligned} &F_D(\mathbf{w}) = \tfrac{\gamma_L}{2}||\mathbf{w}||^2 + \textstyle\sum \xi_{ii'} \\ &s.t.\ y_{ii'}(\mathbf{w}^T x_{ii'} + b) \geq 1 - \xi_{ii'} \end{aligned} \tag{7}$$

where $\xi_{ii'}$ denotes the slack variables that allow the model for nonlinearly 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 model.

6.2 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 7. 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 7. A main strongly-connected cluster formed by correctly linked users (the dashed red arrows in Figure 7) would generate agreement links (edges with positive weights) among one another. These links are formed when behaviors between 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 7). 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 still be constructed and propagated along the social structure even when the ground truth linkage information is not available for every training user pairs.

To model the structure consistency, first, a set of candidate matchings 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 $\mathbf{M}$.

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:

$$\begin{aligned} &\min_{\mathbf{w}} F_S(\mathbf{w}) = \mathbf{w}^T X^T (\mathbf{D} - \mathbf{M}) X \mathbf{w} \\ &s.t.\ ||\mathbf{w}||^2 \le s, D(a, a) = \textstyle\sum_b M(a, b) \end{aligned} \tag{8}$$

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 $c \le c', c = 1, ..., C-1$ 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:

$$\begin{aligned} M^{cc'}(a, b) = &\exp\left(\frac{-\left(||x_i - x_{i'}||^2 + ||x_j - x_{j'}||^2\right)}{2\sigma_1^2}\right) \cdot \\ &\left(1 - \frac{(d_{ij} - d_{i'j'})^2}{\sigma_2^2}\right) \end{aligned} \tag{9}$$

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. 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.

6.3 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 [19]:

$$\begin{aligned} &\min_{\mathbf{w}} F(\mathbf{w}) = [F_D(\mathbf{w}), \mathbf{F}_S(\mathbf{w})] \\ &s.t.\ c_{ii'}(\mathbf{w}^T x_{ii'} + b) \ge 1 - \xi_{ii'}, i \in \mathsf{S}, i' \in \mathsf{S}', ||\mathbf{w}||^2 \le s \end{aligned} \tag{10}$$

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, constraint $||\mathbf{w}||^2 \le s$ can be omitted. Therefore, we define a utility function to aggregate all the objective functions in the form of a generalized weighted exponential sum as:

$$U = \sum_{k=1}^{(C-1)C/2+1} w_k [F_k(\mathbf{w})]^p, \forall k, F_k(\mathbf{w}) > 0, w_k \ge 0 \tag{11}$$

where the weight parameter w_k is a preference parameter. By minimizing utility function U, we seek the Pareto optimal solutions [19], which cannot be improved for 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.* [1] and Yu [30] for details. □

When $p = 1$, this utility function is similar to traditional semi-supervised learning models with a weighted combination of empirical loss, the penalty on $\mathbf{w}$ and a graph Laplacian regularizer [2]. When $p > 1$, our model can be viewed as minimizing the distance function between the solution point and *Utopia* points [1] 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 [25], the decision function $\mathbf{w}$ can be expressed by the dual problem as the expansion over labeled user pairs and unlabeled candidate user pairs $\mathbf{w} = \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \alpha_{ii'} \phi(x_{ii'})$. Then, the

decision function is given by:

$$f(x_t) = \sum_{ii' \in \mathbb{P}_l \bigcup \mathbb{P}_u} \alpha_{ii'} K(x_{ii'}, x_t) + b \tag{12}$$

where we use $\mathbf{K}$ to denote the kernel matrix formed by kernel functions $K\left(x_{ii'}, x_{jj'}\right) = \langle \phi\left(x_{ii'}\right), \phi\left(x_{jj'}\right) \rangle$. 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. 12 into Eqn. 11 and introduce the *Lagrangian* multipliers, and obtain the following regularized utility function to be minimized:

$$\begin{array}{l} \min\limits_{\alpha,\beta} \left\{ \frac{1}{2}\alpha^T \left(2\gamma_L \mathbf{K} + \frac{2\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} \mathbf{K}(\mathbf{D} - \mathbf{M})\mathbf{K} \right) \alpha \right. \\ \left. -\alpha^T \mathbf{K}\mathbf{J}\mathbf{Y}\beta + \beta^T \mathbf{1} \right\} \end{array} \tag{13}$$

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:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}^{12} & 0 & ... & 0 \\ 0 & ... & ... & ... \\ ... & ... & \mathbf{M}^{cc'} & 0 \\ 0 & ... & 0 & ... \end{bmatrix} \tag{14}$$

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:

$$\alpha = \left(2\gamma_L \mathbf{I} + 2\frac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} (\mathbf{D} - \mathbf{M}) \mathbf{K} \right)^{-1} \mathbf{J}^T \mathbf{Y} \beta^* \tag{15}$$

Again, substituting Eqn. 15 into the dual function Eqn. 14, we obtain the following smooth quadratic programming problem to be solved:

$$\begin{array}{l} \beta^* = \max\limits_{\beta} \left\{ \beta^T \mathbf{1} - \frac{1}{2}\beta^T \mathbf{Q} \beta \right\} \\ s.t. \sum\limits_{ii' \in_l} \beta_{ii'} y_{ii'} = 0, 0 \leq \beta_{ii'} \leq \frac{1}{|\mathbb{P}_l|} \end{array} \tag{16}$$

where:

$$\mathbf{Q} = \mathbf{Y}\mathbf{J}\mathbf{K} \left(2\gamma_L \mathbf{I} + 2\frac{\gamma_M}{|\mathbb{P}_l \bigcup \mathbb{P}_u|^2} (\mathbf{D} - \mathbf{M}) \right)^{-1} \mathbf{J}^T \mathbf{Y} \tag{17}$$

The derivation above shows that the entire user identity linkage problem can be well cast into a standard convex programming problem that can be solved by many off-the-shelf optimization packages. Although we only introduce the model construction for $p = 1$, similar derivation can also be readily performed for $p > 1$. The resulting objective function is also convex due to the convexity of the individual objective functions and the convexity of the utility function U. Moreover, setting higher values for p would increase the effectiveness of the method in providing the complete Pareto optimal set [1, 19].

Dealing with Missing Information. When constructing the pair-wise similarity between users, it is not uncommon to observe a significant amount of information missing among real-life social platforms due to: (I) intrinsically heterogeneous information sources, (II) unpredictable user log-in and log-out, and (III) privacy concerns. Previous approaches [16, 32] construct discriminate models where a missing feature is automatically filled with zeros based on the assumption that the values do exist but are not observed, which, unfortunately, is not the case for our problem. To effectively handle missing information, we fill the missing information by making use of the core social network structure. For each user pair, we denote their top-3 interacting friends as i_1, i_2, i_3, and i'_1, i'_2, i'_3. The average behavior similarity and the standard deviation of the social connection of user i and i' can be calculated as:

$$\overline{s}(i, i') = \frac{\sum_{p=1}^{3} \sum_{q=1}^{3} s(i_p, i'_q)}{9} \tag{18}$$

where $s(i_p, j_q)$ denotes the similarity of any particular similarity measure as described in previous sections. If the information of their friends are still missing, we automatically fill the corresponding dimension as 0.

Due to the extremely large data size, we adopt the distributed convex optimization method [3] 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.

6.4 Model Analysis

We learn the linkage function via optimizing two kinds of objective functions, i.e., the supervised learning using the reliable ground truth, and the structure consistency maximization by modeling the core social network behavior consistency. They are complementary to each other by jointly measuring user behavior similarity at both individual and group levels. When the ground truth information is insufficient (e.g., less than 10% of the pairs are assigned with labels), the model will be more dependent on the core social network structure. The linked user pairs will be served as "*anchor*" pairs from which the linkage information can be propagated along the core social network. 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 are assigned with labels), the model can still be endowed with greater generalization power by the decision boundary smoothed towards better group level behavior consistency. The l_p-norm in the utility function determines how the two kinds of objective functions interact with each other, such that a larger p imposes greater uniqueness on the dominant objective function. Correspondingly, model over-fitting 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.

Algorithm 1 The HYDRA algorithm

Input: Data: X,Y, Parameters: γ_L, γ_M, p, σ_S, σ_D
Output: α, β

1: Select the candidate pair set $\mathbb{P}_u$ by comparing the pair-wise similarity.
2: Construct structure consistency graph $\mathbf{M}$.
3: **while** the stopping criterion is not reached **do**
4: Find optimal β^t by solving (Eqn. 16).
5: **end while**
6: Obtain α^t by Eqn. 15.

7. EXPERIMENTAL EVALUATION

7.1 Experiment Setup

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 network service dubbed as the Facebook of China with 162 million registered users.
4. **Douban:** (www.douban.com) A social network 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 network service with 160 million registered users.

We use 5 million Chinese 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 Chinese 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 social networks above, 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), and 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 result, x-axis is the 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 [32].

(II) Alias-Disamb: an unsupervised data-driven approach based on username analysis to link users across platforms [16].

(III) SMaSh: a record linkage approach finding linkage points over Web data [11].

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

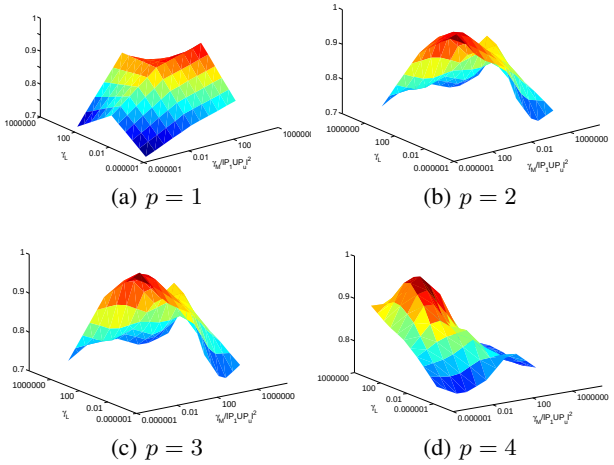
(V) HYDRA-Z: a degenerate version of our model HYDRA where all the missing features are filled with zeros.

(VI) HYDRA-M: our model HYDRA with missing features filled with the core social network friend structure described in Section 7. 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.

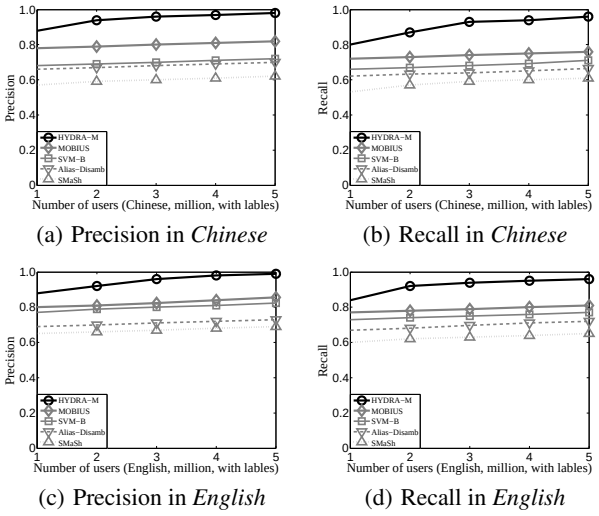
For the pair-wise similarity calculation in this paper, the parameters (e.g., ε for user profiling, q and λ for multi-resolution temporal similarity modeling) are tuned by a grid search procedure to maximize the performance of a linear SVM on the validation set. Then the optimized multi-dimensional similarity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

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.



(a) $p = 1$ (b) $p = 2$ (c) $p = 3$ (d) $p = 4$

Figure 8: Performance curve with various settings of γ_M and γ_T under varied p.



(a) Precision in *Chinese* (b) Recall in *Chinese*

(c) Precision in *English* (d) Recall in *English*

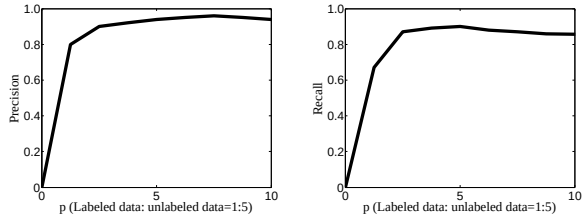
Figure 9: Performance w.r.t. #labeled pairs.

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.

The parameters of all the kernels for HYDRA are tuned strictly according to the methods described in the previous sections.

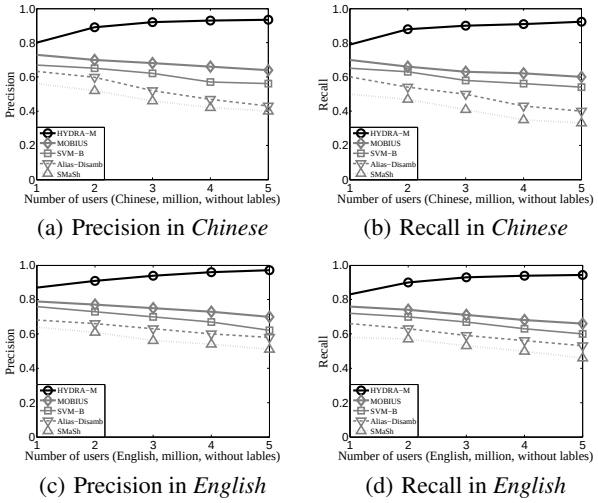
7.2 Effectiveness Evaluation

Performance w.r.t. Varied 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 8. From Section 6 we see that γ_M and γ_L determine the relative importance of the problems in MOO framework from a decision



(a) Precision (b) Recall

Figure 10: The precision and recall curve w.r.t. different p.



(a) Precision in *Chinese* (b) Recall in *Chinese*

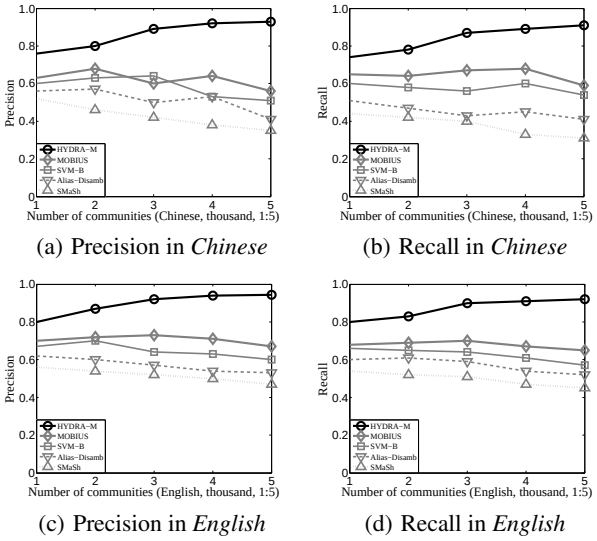
(c) Precision in *English* (d) Recall in *English*

Figure 11: Performance w.r.t. #unlabeled pairs.

maker's perspective, while p determines how the learned model approximates the *Utopia* solution, thus determining the intrinsic structure of the utility function. However, for real data, a decision maker's preference does not necessarily give the best performance, as can be seen from Figure 8. The results show that different settings of p lead to different optimal settings of γ_M and γ_L.

Performance w.r.t. Varied p. Figure 10 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 over-fitting. We see from Figure 10 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. Varied Numbers 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 experiment results are reported in Figure 9. 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 that on Chinese ones, which is also true for Figure 11. 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, the social



(a) Precision in *Chinese* (b) Recall in *Chinese*

(c) Precision in *English* (d) Recall in *English*

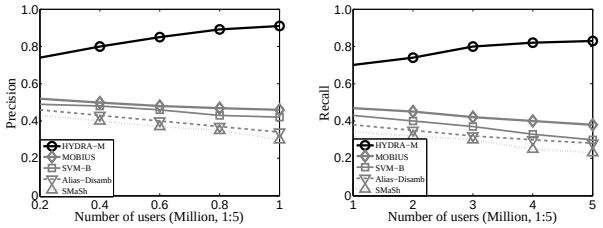
Figure 12: Performance with #social communities.

structure and user behavior on Chinese platforms are characterized with greater complexity and a higher level of temporal dynamics than those on English platforms. Taking Twitter and Sina Weibo for example. In comparison, Sina Weibo has much more retweets and a greater diffusion speed for retweet than Twitter, which means the information diffusion in Sina Weibo is much faster than in Twitter. We also find out that Sina Weibo contains much richer and more dynamic information than Twitter, presenting a much more challenging task for the *SIL* problem. Note that most users in Sina Weibo have much more followers and followees than those in Twitter. Consequently, the much more complicated social structure contributes to the greater challenge for the *SIL* problem on Chinese platforms.

Performance w.r.t. Varied 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 11. Compared against Figure 9, we notice that the performance of baseline methods with unlabeled data is much worse than the performance with labeled data in Figure 9. But our HYDRA survives the unlabeled data setup and performs much better than the baseline methods. In Figure 9 and Figure 11, HYDRA not only performs much better (higher precision and recall) than the baseline methods, but also shows better performance along with the increasing number of users.

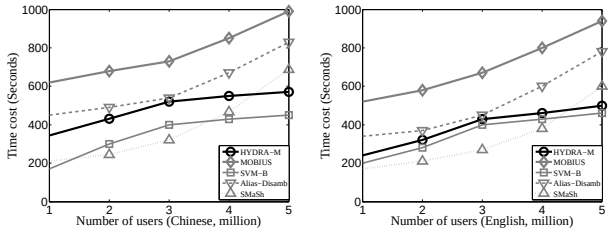
Performance w.r.t. Varied Numbers of Social Communities. We evaluate how the structure information from other social communities [6] could help enhance the model generalization power. Specifically, given the top five largest overlapping communities A, B, C, D, E with labeled training pairs between A 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 12. An interesting observation is that 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. But as we notice in Figure 12, social community structure indeed help HYDRA achieve better results than baseline methods.

Performance w.r.t. Varied Social Platforms. We study *SIL* across



(a) Precision (b) Recall

Figure 13: Performance on various social platforms.



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

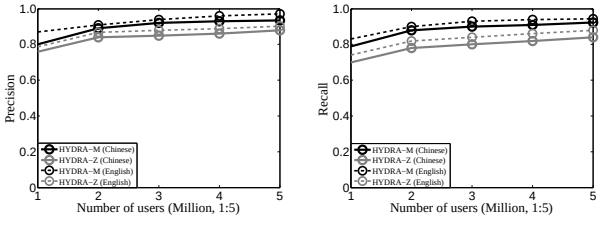
Figure 14: Efficiency evaluation.

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 13. 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 by varying (I) parameter settings, (II) numbers of labeled data pairs, (III) structure information levels and (IV) numbers of social communities, we conclude that HYDRA significantly outperforms the baseline methods and displays good scalability with the increasing amount of data, for both Chinese and English platforms.

7.3 Efficiency Evaluation

We use the total execution time at different scales to evaluate the efficiency. From the results reported in Figure 14, HYDRA consumes less time than the baseline methods (except SVM-B and SMaSh) on 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 social structure we incorporate into the HYDRA model — for such a five-million-user social network, when we have accumulated around three million users and their one-hop friends, the social structure is almost well-constructed. Then the resulting utility function would contain 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



(a) Precision (b) Recall

Figure 15: Performance with missing data.

[16], 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*. In conclusion, HYDRA is capable of handling large-scale data sets.

7.4 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 15, 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.

7.5 Discussion

The data sizes in this paper are prohibitively large for a single PC or server. Despite that we deal with millions of user records when optimizing the convex problem in Eqn. 16, the problem can still be handled efficiently by several servers for the following reasons.

Firstly, their behavior information are extremely sparse. For example, the total amount of non-missing or non-zero features on the data of English communities are no more that 4%, and the amount of available similarities between users are no more than 2%. Even though some missing values can be filled using the core social structure, the amount of available similarities are still no more than 3%. Similarly, the available similarities are about 2% on the data of Chinese communities. Besides, the structure consistency matrix $\mathbf{M}$ is even more sparse, which typically contains less than 1% non-zero elements for both English and Chinese communities. Such data sparsity allows efficient data storage and successful execution of our learning algorithm with 5 high-end servers in our experiments.

Moreover, we learn the model by a distributed optimization method [3] 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 the optimization of a series of sub-problems on different parts of the data stored distributively across different servers. Meanwhile, our model uses support vector representation, i.e., α and β, where at least 90% of the dimensions in β are zeros on a million-scale data. For each step of the model optimization, we perform a coefficient space shrinking process to actively identify the non-zero dimensions in β with

a simple gradient thresholding technique. Consequently, the corresponding zero entries 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 improve the efficiency of the model learning by using β^t as a warm start to optimize β^{t+1}.

8. CONCLUSION

In this paper, we link up user accounts of the same natural person across different social network platforms. We propose a framework, HYDRA, a multi-objective learning framework incorporating heterogeneous behavior modelingvand core social network structure. We evaluate HYDRA against the state-of-the-art solutions on two real data sets — five popular Chinese social networks and two popular English social networks, a total of 10 million users and more than 10 tera-bytes of data. Experimental results demonstrate that HYDRA outperforms existing algorithms in identifying true user linkage across different platforms.

9. ACKNOWLEDGMENTS

This research has been 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 has also been supported in part by National Basic Research Program of China (973 Program): 2012CB316400, and in part by National Natural Science Foundation of China: 61303160.

10. REFERENCES

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

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

[3] 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*, 3(1):1–122, 2011.

[4] J. Cai and M. Strube. End-to-end coreference resolution via hypergraph partitioning. In *COLING'10*.

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

[6] W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang. Online search of overlapping communities. In *SIGMOD'13*.

[7] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4):55–64, 2001.

[8] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–16, 2007.

[9] R. C. Gonzalez and R. E. Woods. Digital image processing. 1992.

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

[11] 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*, 6(6):444–456, 2013.

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

[13] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *ICWSM'11*.

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

[15] S. Kumar, R. Zafarani, and H. Liu. Understanding user migration patterns in social media. In *AAAI'11*.

[16] 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*.

[17] S. Liu, S. Wang, K. Jeyarajah, A. Misra, and R. Krishnan. TODMIS: Mining communities from trajectories. In *ACM CIKM'13*.

[18] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida. Studying user footprints in different online social networks. In *ASONAM'12*.

[19] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, 26(6):369–395, 2004.

[20] Y. nan Qian, Y. Hu, J. Cui, Q. Zheng, and Z. Nie. Combining machine learning and human judgment in author disambiguation. In *CIKM'11*.

[21] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In *WWW'04*.

[22] A. Nunes, P. Calado, and B. Martins. Resolving user identities over social networks through supervised learning and rich similarity features. In *SAC'12*.

[23] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[24] M. Sadinle and S. E. Fienberg. A generalized fellegi-sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 105(502):385–397, 2013.

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

[26] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD '13*.

[27] J. Vosecky, D. Hong, and V. Shen. User identification across multiple social networks. In *NDT'09*.

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

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

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

[31] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *ICWSM'09*.

[32] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *KDD'13*.

[33] 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*, 57(3), 2006.