Discovering Multi-Relational Structure in Social Media Streams

YU-RU LIN
Arizona State University

HARI SUNDARAM
Arizona State University

MUNMUN DE CHOUDHURY
Arizona State University

AISLING KELLIHER
Arizona State University

In this article, we present a novel algorithm to discover multi-relational structures from social media streams. A media item such as a photograph exists as part of a meaningful inter-relationship amongst several attributes including – time, visual content, users, and actions. Discovery of such relational structures enables us to understand the semantics of human activity and has applications in content organization, recommendation algorithms, and exploratory social network analysis.

We are proposing a novel non-negative matrix factorization framework to characterize relational structures of group photo streams. The factorization incorporates image content features and contextual information. The idea is to consider a cluster as having similar relational patterns – each cluster consists of photos relating to similar content or context. Relations represent different aspects of the photo stream data, including visual content, associated tags, photo owners, and post times. The extracted structures minimize the mutual information of the predicted joint distribution. We also introduce a relational modularity function to determine the structure cost penalty, and hence determine the number of clusters. Extensive experiments on a large Flickr dataset suggest that our approach is able to extract meaningful relational patterns from group photo streams. We evaluate the utility of the discovered structures through a tag prediction task and through a user study. Our results show that our method based on relational structures, outperforms baseline methods, including feature and tag frequency based techniques, by 35%–420%. We have conducted a qualitative user study to evaluate the benefits of our framework in exploring group photo streams. The study indicates that users found the extracted clustering results clearly represent major themes in a group; the clustering results not only reflect how users describe the group data but often lead the users to discover the evolution of the group activity.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: *Information Search and Retrieval*---Information filtering; H.3.5 [**Information Storage and Retrieval**]: *Online Information Services*---Web-based services; H.4.3 [**Information Systems Applications**]: *Communications Applications*; H.5.1 [**Information Interfaces and Representation**]: *Multimedia Information Systems*---Evaluation/methodology; H.5.4 [**Information Interfaces and Representation**]: *Hypertext/Hypermedia*
General Terms: Experimentation, Measurement, Algorithms, Human Factors
Additional Key Words and Phrases: Social media, Social network analysis, Structure mining, Multi-relational learning, Nonnegative matrix factorization

---

1. INTRODUCTION

In this paper, we present a novel algorithm to discover multi-relational structures from social media streams. In particular, we discover latent structures in the popular online social media site, Flickr. The structure encodes relational semantics pertaining to human activity within such social networks. This latent structure has the potential to enhance content organization, improve recommendation algorithms, and support exploratory social network analysis.

The semantics of human activity on social media sites (including Flickr and YouTube) needs to be understood as a relationship between people, actions, artifacts, and supportive contextual metadata. Today, social media sites have made it easy to upload, share, and interact with content as well as to communicate with other users on the site. Each site

provides a diverse range of functionalities, including tagging and commenting on media, as well as direct communication with other users. A shared media item is associated with a variety of information, including the identity of the person who uploaded it, associated tags, identities of people who commented on the item, and the number of times viewed or marked as a "favorite". These user actions (including "upload," "tag," "comment") are additionally associated with a timestamp. A media item such as a photograph on Flickr therefore exists as part of a meaningful inter-relationship amongst several attributes including time, visual content, users, and actions (Figure 1[1]).

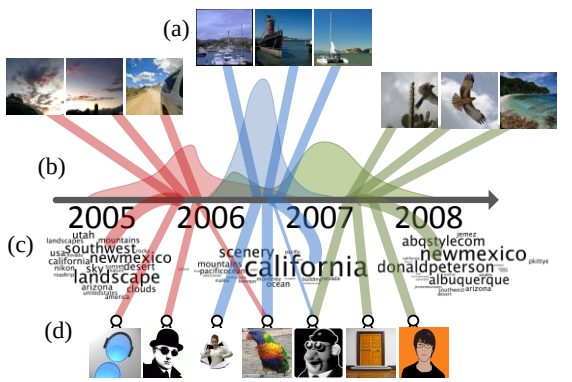

Figure 1: Relational structure in social media streams, which reveals the strong relationship among multiple facets such as (a) photos (visual content), (b) time, (c) tags and (d) users. This figure presents partial multi-relational structure extracted from the data of Flickr group "The Southwest United States", based on our proposed algorithm. More detailed results can be referred to section 6.2. It illustrates three major themes in the group photo stream: "landscape," "california" and "newmexico," which mostly associate with three time frames (2005-2006, 2006-2007 and 2007-2008), where different users contributed photos to these themes.

Relational semantics derived from human activity are distinctly different from media semantics (e.g. "what is the meaning of this photo?"). For example, a Flickr group on "Arizona Travel" may have a lot of posts on Sedona, a popular destination, in July from people who live in Phoenix who travel there to escape the heat. There are fewer posts in December, when it is cold in Sedona. Now, the meaning of the relationship between location (Sedona), time (summer), specific users, and photo colors is not explicit in the data. This relationship may exist because the active members of the group are friends who live in Phoenix, and plan an annual summer retreat together in Sedona. In this case, the relational semantics, while not explicit, are known only to the group members. These relational semantics cannot be easily discovered by accessing the photo stream via a single object or attribute (e.g. photo tags), or through a simple aggregation of attributes. Interestingly, the discovery of relational structure in such social media sites can point to emergent cultural behaviors, which may not even be explicitly identifiable by members of the network.

1.1 Motivating Applications

We discuss how relational pattern discovery can significantly impact content organization, recommendation algorithms, and exploratory social network analysis.

Content organization. The rapid growth of content on social media sites creates several interesting challenges. First, the content in a photo stream (either for a user or a community) is typically organized using a temporal order, making the exploration and browsing of content cumbersome. Second, sites including Flickr provide frequency based aggregate statistics including popular tags and top contributors. Users can access a subset of the content by clicking on these tags / contributors. However, these aggregates do not reveal the rich relational structure inherent in the community sharing and interaction. As discussed in [Shamma et al. 2007], how to harness the contextual information for media understanding is one of the most difficult challenges in media pragmatics/applications.

Recommendation algorithms. The multi-relational structure can be used to provide effective recommendations along any attribute. When the user is looking at a particular

[1] The original images from the groups are unavailable for copyright reasons. In this article, we have included copyright-free images, similar in appearance to the representative group images.

photo, we could use the set of relations in which the photo is determined to exist, and then recommend other photos, tags, and related peers. The multi-relational data can provide additional context, over the (photo, tag) pairs that have been used to recommend tags in automated annotation algorithms. It can help identify peers and context (including feature distributions, activities, time) in which they are related to the current user.

Social network analysis. An important motivation of this work is to discover an interpretable structure in social media streams to facilitate exploratory analysis. How do groups emerge in online social networks? Are there specific people, and contexts that explain their emergence? Within such groups, how does information flow? What roles (including "aggregators", "disseminators") do users play? Finally, when do new associations between tags and photos emerge? We believe that relational structures can facilitate an effective exploration and summarization of social media.

1.2 Our Approach

A key contribution of this work is the idea that the *semantics* of human activity in such rich social media sites can be understood though the *relational structure* latent in such social networks. The structure encodes the relationship between users, artifacts, and context. It is important to emphasize that the relational semantics are often only available to the participants of the social network. That is, the *explanation* for the existence of a stable relationship between a specific set of people, location, time, photos, and tags, while known to the users, may not be explicitly encoded in the data. Hence, when members of the network are exposed to the latent relational semantics, they may be able to use this information more effectively than non-group members.

We extract the relational structure through a novel algorithm using data from Flickr group photo streams. We define a *group photo stream* (or group for short) to be a collection of photos posted in a social media *sharing space*, together with all the users (who posted the photos) and tags associated with the photos. In this work, the sharing space specifically refers to Flickr group pools (http://www.flickr.com/groups/), while other types of sharing spaces such as Facebook groups can also be considered. Our goal is to extract structures of these relations within the groups by finding a *soft clustering* structure that reflects these relations, which we call *relational clusters*, where each data item (e.g. a photo, tag, etc.) is assigned to multiple clusters with membership weights that sum to one.

There are three key ideas in our framework:

(1) *Extraction of relations in social media streams*: Relations represent different aspects of the photo stream data, including visual content, associated tags, photo owners, and post times. We consider such content and contextual information as various relations. The relational representation allows analyzing the correlation across different aspects (as in our joint factorization described below) as well as easily incorporating richer contextual information.

(2) *Extraction of relational clusters*: We formulate the extraction of relational clusters as an optimization problem where the objective is to find a set of soft clusters that best represent our observation of various relations simultaneously. We incorporate various relations in a non-negative joint matrix factorization framework and solve the objective by a scalable iterative algorithm with linear time complexity.

(3) *Evaluation of structure complexity*: We propose a relational modularity function for evaluating the clustering structure and determining the optimal number of relational clusters. We show that the quality of a clustering defined by the factorization objective can be interpreted as maximizing the mutual information of the predicted joint distribution. The relational modularity is then defined as a function of the quality and cost of the structure, where the cost is determined by the mutual information conditioned on the clustering. We also show a close connection between

the relational modularity function and the modularity concept first introduced in [Newman and Girvan 2004].

We have conducted extensive experiments on large, real world Flickr datasets that include 120 Flickr photo groups, with 111,108 photos posted over five years. We present case studies on the structures extracted from four Flickr groups. The analysis reveals consistent clustering structures in each group in terms of time profiles, visually coherent photos, and blocking structures in relational matrices. We compare the structures extracted from these Flickr groups with random groups (collections of randomly selected photos) and observe that the structures of random groups are qualitatively different from the relational structures extracted from collective human activities. We evaluate the utility of the discovered structures through a tag prediction task. Our prediction results outperform baseline methods including feature and tag frequency based techniques, by 35%–420% (based on NDCG) on an average. The results suggest that our analysis based on relational clustering structures helps improve the quality of tag prediction and provides a quantitative evaluation for the meaningfulness of an extracted structure. We have conducted a pilot user study with 12 participants to understand the impact of exposing the group relational semantics to them. The study indicates that users found the extracted clustering results clearly represent major themes in a group; the clustering results not only reflect how users describe the group data but often lead the users to discover the evolution of the group activity.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 discusses relations in social media streams. Section 4 and 5 present our method for extracting relational clusters and for evaluating the structure complexity. Section 6 reports our experimental study. Section 7 discusses the open issues; section 8 presents our conclusions. The appendix contains proofs, and detailed user study results.

2. RELATED WORK

In this section, we first briefly discuss applications based on structure mining in different domains, and then situate our work within the context of online social media analysis. We further discuss related techniques for analyzing multimodality or multi-relational data.

Structure mining. Mining structures from data arise naturally in many applications. Xie et al. [2002] propose a hidden Markov models (HMMs) based approach for mining temporal structures in video streams. The structures of interest are repetitive segments that often relate to recurrent events, e.g., plays and breaks in a soccer video. Mccowan et al. [2005] propose a framework for extracting the structure of group meeting actions such as monologue, discussion, note-taking, etc. from audio-visual streams. “Topic models” (e.g. [Blei et al. 2003]) discover patterns in text corpus by representing the underlying topics with word distributions and the topics are combined to form documents. In addition, dynamic topic models [Blei and Lafferty 2006; Wang and Mccallum 2006] are developed to capture the evolution of topics in a sequentially organized corpus of documents. In social network analysis, the structures of interest do not necessarily correspond to explicit semantics and must be interpreted according to the time and context of the observation (e.g. US Supreme Court rulings [Doreian and Fujimoto 2001]). Such analysis has provided important insights for social functions and processes. In this paper, we are interested in relational structures that deviate from the structure mining within multimedia, moving more towards the goals of structure mining within social networks, which may have emergent or context-bounded semantics. Our idea can be stated in a way analogous to the topic modeling methods – we consider the relational structures as coherent distributions of different types of social media features (e.g. users, tags, visual features, etc.) which combine to form the social media streams.

Social media analysis. The popularity of social applications and environments has attracted considerable research interests. In particular, our work relates to two primary

directions – namely, studying people's social networking behavior, and improving the search and recommendation of media by exploiting contextual as well as social knowledge. The first direction is often studied based on the statistical properties of online social networks. Backstrom et al. [2006] study how the structural features correlate with changes of social group members. Kumar et al. [2006] study the evolution of the blogosphere in terms of the change of graph characteristics and the community burstiness, where a temporal burst is defined based on hyper-linking occurrences. Palla et al. [2007] extract groups per time slice and then quantifies their evolution based on membership differences. The structure of social interactions among people have also been studied through a unipartite or bipartite graph, in which the community structure can be characterized by clustering methods [Lin et al. 2008; Sun et al. 2007].

The second direction considers utilizing the media metadata associated with media objects to improve media content retrieval. Existing research on tagging services includes improving tag recommendation [Garg and Weber 2008; Sigurbjörnsson and Van Zwol 2008], and analyzing usage patterns of tagging systems [Negoescu and Gatica-Perez 2008]. Chen et al. [2008] propose a group and tag recommendation system by using concept detectors trained based on visual features and tags. Negoescu and Gatica-Perez [2008] present a large scale analysis of Flickr groups and propose a topic modeling approach for representing a group based on the co-occurrence of groups and tags. Zunjarward et al. [2007] propose a framework for annotating events in images by exploiting the social networks of annotators. Kennedy et al. [2007] propose a framework for generating knowledge (representative tags) for a location, and for extracting place and event semantics for a tag. Their work suggests that community-generated media and tags can improve access to multimedia resources. Similarly, Ahern et al. [2007] propose a system to analyze the tags associated with the geo-referenced images to generate knowledge about a given location in the form of representative tags. These tags can be further utilized to help expose the photo content. Shamma et al. [2007] design and prototype tools for capturing the context in which the media is used, and investigate methods for using the information to organize and index media.

Multi-relational learning. As discussed in [Li and Anand 2007], classical propositional clustering methods are limited in dealing with data having various types of entities and different semantic relationships among them. The first-order extensions of classical methods (e.g. RDBC [Kirsten and Wrobel 1998]) may not be feasible for large relational data due to the quadratic computational complexity. Recently, relational clustering techniques have been proposed to learn the interrelated structures among various entities and multiple relationships [Banerjee et al. 2007; Bekkerman et al. 2005; Long et al. 2006; Tang et al. 2008; Wang et al. 2006]. However, the implementations of these techniques can only handle small-scale datasets and do not take advantage of the sparseness in social network data. In addition, most of the relational clustering techniques consider hard clustering assignment (i.e. each entity can belong to only one cluster), which limits its direct usage in applications such as context-sensitive recommendation in social media.

In multimedia, techniques for combining multimodality in media content analysis have been developed to overcome the semantic gap problem since visual features are unable to represent the image content at the semantic level. Cai et al. [2004] use visual, text and link information of images to construct a relationship graph of Web images. Tong et al. [2005] propose a graph based learning approach (both semi-supervised and unsupervised), where different modalities are represented independently as a multiple graphs. Rege et al. [2008] propose a graph theoretical framework for simultaneously integrating visual and textual features for co-clustering a tripartite (feature-image-word) graph. Multi-graph mining has also been studied in other contexts. In text mining, Zhu et

al. [2007] propose a matrix factorization algorithm combining both the linkage and the document-term matrices to improve the hypertext classification.

This article extends our prior work [Lin et al. 2009a] – we include formal discussion of the problem, extended solutions, detailed algorithms, proof, and new experiment results. Another related work is [Lin et al. 2009b] which proposes a framework called JAM for constructing a user-centric summary of their online social activities. Instead of a fixed syntactical structure for representing user activities, i.e. correlating users and concepts through different actions, this work considers a generalized relational data model to represent data in social media streams. In addition, the JAM framework extracts one dominate theme at each time, while in this work, multiple relational clusters can co-exist, which allows users to compare or contrast multiple themes.

Our unique contribution. Mining time-evolving relational patterns in social media streams deals with the interrelatedness of media content features as well as contextual and temporal information associated with the media. Our analysis aims to generate an interpretable structural representation of the social media stream and allow the flexible incorporating of various media relations (visual, temporal, etc.). The structural representation with probabilistic interpretation can be used directly to retrieve different types of representative entities and to summarize different aspects of a social media stream. Our method also takes advantage of the sparseness in social network data, which is able to handle large scale data in social media streams.

3. RELATIONS IN SOCIAL MEDIA STREAMS

In this section, we discuss various relations often observed in community generated media environments. We first introduce the concept of relations. Then we specifically discuss relations within the Flickr group photo streams, including visual features (section 3.1) and the photo information context (section 3.2).

Let us assume we observe a set of photos. These photos are posted by a set of users at certain times, associated with a set of tags, and consist of a set of visual features. We call a set of objects or entities of the same type a *facet*, e.g. a photo facet is a set of photos, a user facet is a set of users, etc. We call the interactions among facets a *relation*; a relation can involve two (i.e. binary relation) or more facets. In this work we only investigate pairwise relations, but our method can be extended to higher ordered relations. We discuss the extraction of specific relations in the following subsections.

3.1 Visual Features

We use a number of image features that have been found effective in image content analysis. We briefly summarize these features as follows:

(1) **Color**: We use two color based features, color histogram and color moments.
(2) **Texture**: We use a phase symmetry [Xiao et al. 2005] method for detecting textures from arbitrary "blobs" in images. It is based on determining local symmetry and asymmetry across an image using phase information.
(3) **Shape**: We use two shape features, radial symmetry [Loy and Zelinsky 2003] and phase congruency [Liu and Laganiere 2007]. The radial symmetry feature detects points of interest in an image. Phase congruency is an illumination and contrast invariant measure of feature significance.
(4) **Interest Points**: We extract local interest point descriptors for a image by scale-invariant feature transform (SIFT) [Lowe 2004]. SIFT features have received much interest due to their invariance to scale and rotation transforms and their robustness against changes in viewpoint and illumination.

After extracting these features, we construct a D-dimensional feature vector for each photo, where D=1064 in this work. Let P be the set of group photos, we obtain a photo-feature matrix $\mathbf{W}^{(F)} \in \Re^{|P| \times D}$, where the i-th row is the feature vector of the i-th photo. We

use standard normalization with a logistic function $g(x)=1/(1+\exp(-x))$ to bound the feature values within the range [0,1]. This matrix $\mathbf{W}^{(F)}$ represents the relation between the photo and the visual features.

3.2 Contextual Information

We now discuss the context associated with a photo. In social media, a media object such as a photo generally has rich contextual information, e.g. who shares the photo, when the photo is shared, and what additional information is associated with the photo. Three kinds of basic contextual information are discussed as follows:

(1) **Users**: The content and concepts of a photo are determined by its owner, and hence the ownership provides the most important contextual information. Let U be the set of users who post photos to the group, i.e. photo owners. We construct a photo-user matrix $\mathbf{W}^{(U)} \in \Re^{|P| \times |U|}$, where each entry $\mathbf{W}_{ij}^{(U)} = 1$ if the i-th photo is posted by the j-th user, and 0 otherwise.

(2) **Tag**: Tags are descriptive labels assigned by users to describe the content of a photo, e.g. "sky", "bird", etc., or to provide additional contextual and semantic information, e.g. "summer", "vacation", "Nikon", etc. Let Q be the set of tags associated with the group photos. We construct a photo-tag matrix $\mathbf{W}^{(Q)} \in \Re^{|P| \times |Q|}$, where each entry $\mathbf{W}_{ij}^{(Q)} = 1$ if the i-th photo has the j-th tag, and 0 otherwise.

(3) **Time**: A photo in Flickr has a timestamp indicating when the photo was taken or posted (uploaded). Here we use the photo post time since the photo taken time may not be available or may not be set correctly for some photos. We segment the timestamps into S time slots and construct a photo-time matrix $\mathbf{W}^{(T)} \in \Re^{|P| \times S}$, where each entry $\mathbf{W}_{ij}^{(T)} = 1$ if the i-th photo is posted during the j-th time slot, and 0 otherwise. We smooth the temporal information by applying a symmetric moving average filter on each row vector of $\mathbf{W}^{(T)}$:

$$y_i = \frac{1}{M} \sum_{j=-(M-1)/2}^{(M-1)/2} b_j x_{i+j} ,$$

where x is the input signal, y is the output signal, M is the filter size and b_j's are the filter coefficients. A filter kernel is determined by the filter coefficients. E.g., two different kernels, rectangular and triangular, with filter size M=5 are shown in Figure 2. Depending on the length of a time slot, we can choose to use either type of filter (usually the rectangular kernel is used for short-length time slots). With such smoothing filter, two photos associated with nearby time slots become similar in the temporal dimension. In other words, the smoothing filter is used to propagate the "temporal similarity" to nearby time slots.

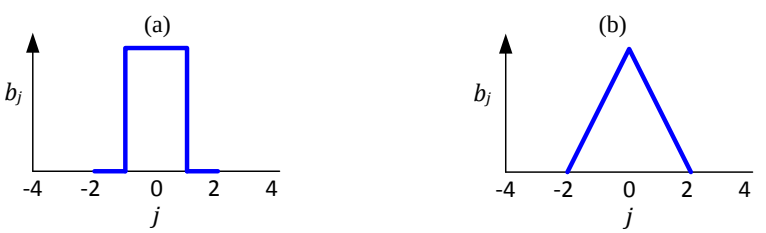

Figure 2: The temporal information in photo-time matrix $\mathbf{W}^{(T)}$ is smoothed by a moving average filter with, e.g. filter size M=5, and (a) rectangular kernel or (b) triangular kernel. Depending on the length of a time slot, we can choose to use either type of filter (usually the rectangular kernel is used for short-length time slots). When applying a smoothing filter on the photo-time matrix, two photos associated with nearby time slots become similar in the temporal dimension.

In our analysis, the data representing the group photo streams over time is given as the data matrices described above, including the image content features as well as the image context and temporal information. Each relation is represented by a matrix. In our notations, a matrix $\mathbf{W}^{(\cdot)}$ is indexed based on its second dimension. Without loss of

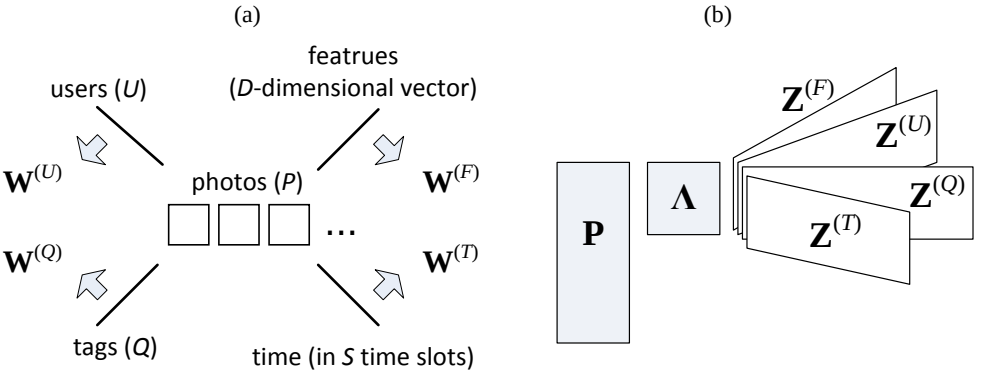

Figure 3: (a) Data representation: The relations in a group photo stream is given as four matrices: photo-feature matrix $\mathbf{W}^{(F)}$, photo-user matrix $\mathbf{W}^{(U)}$, photo-tag matrix $\mathbf{W}^{(Q)}$ and photo-time matrix $\mathbf{W}^{(T)}$. The first matrix comprises visual information and the last three matrices comprise contextual information of the photo stream. (b) We use joint factorization to extract soft clusters from various relations simultaneously. The relational data represented as in Figure 3(a) are factorized by the photo-cluster matrix **P** and a set of coefficient matrices for different facets (visual features, users, tags, and times).

generality, we normalize $\mathbf{W}^{(\cdot)}$ to ensure $\Sigma_{ij}\mathbf{W}_{ij}^{(\cdot)}=1$. The data representation is illustrated as in Figure 3(a).

Note that the relational data model based on matrix representation can easily incorporate more contextual information. For example, it is possible to retrieve the EXIF metadata of images and further obtain the image taken time, location, camera settings, etc. Such information can be represented similarly to the basic contextual information discussed in this section. (In this work we do not use EXIF information, due to the overhead of additional API calls for retrieving EXIF metadata for all images.)

4. RELATIONAL CLUSTER EXTRACTION

We now present our method for extracting relational clusters from a group photo stream. We formulate it as an optimization problem (section 4.1), and provide an efficient algorithm to solve the clustering objective (section 4.2).

4.1 Problem Formulation

We formulate the extraction of relational clusters as an optimization problem where the objective is to find a set of *soft clusters* that best represents our observation of various simultaneous relations between the photos and other facets, including: visual features, associated tags, photo owners, and post times. In soft clustering we assume that an entity (e.g. a photo, tag, user, etc.) can belong to multiple clusters, with membership weights that sum to one, indicating how likely the entity belongs to those clusters. We seek to determine soft clustering assignment of entities so that the diverse relations of any two entities can be approximated by their relationship (soft assignment) with a small set of clusters. Each relation is represented by the matrices discussed in the previous section. We now discuss how to extract soft clusters that reflect a specific relation, and then propose a generalized framework for extracting clusters from multiple relations.

Visual features. In order to extract clusters having similar visual content, let us assume each cluster k has a length D feature vector $\mathbf{z}_k$ where each entry z_{kj} can best represent the significance of the j-th feature of the set of photos in the cluster k. Our goal is to determine the coefficient z_{kj} based on how likely a photo i belongs to the cluster k. We define p_{ik} to be the probability that a particular photo i belongs to the cluster k and λ_k to be the cluster probability. The parameters p_{ik} and λ_k are non-negative numbers satisfying $\Sigma_i\, p_{ik}=1$, $\Sigma_k\, \lambda_k=1$. Let $\mathbf{Z}^{(F)}=\{z_{ki}\}$ denote a $K\times D$ matrix, $\mathbf{P}=\{p_{ik}\}$ denote a $|P|\times K$ matrix, and $\mathbf{\Lambda}=\{\lambda_k\}$ be a $K\times K$ diagonal matrix where $\mathbf{\Lambda}_{ij}=\lambda_k$ if $i=j=k$ and 0 otherwise. For

brevity we shall write $\mathbf{\Lambda}_{kk}$ as $\mathbf{\Lambda}_k$. We use an idea similar to principal component analysis to derive $\mathbf{P}$, $\mathbf{\Lambda}$ and $\mathbf{Z}^{(F)}$ from the photo-feature matrix $\mathbf{W}^{(F)}$ (ref. section 3) as:

$$\mathbf{W}_{ij}^{(F)} \approx \sum_k \lambda_k p_{ik} z_{kj} = (\mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(F)})_{ij} \quad <1>$$

This suggests the approximation can be done by minimizing $D(\mathbf{W}^{(F)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(F)})$, given $D(\cdot||\cdot)$ as a measure of approximation cost between two matrices. We use Kullback-Leibler (KL) divergence between two matrices, where the KL divergence is used as a natural measure of the dissimilarity between two distributions. Using matrices to represent distributions, the KL divergence between matrices $\mathbf{A}$ and $\mathbf{B}$ is defined by $D(\mathbf{A}||\mathbf{B}) = \sum_{ij} (\mathbf{A}_{ij} \log \mathbf{A}_{ij}/\mathbf{B}_{ij} - \mathbf{A}_{ij} + \mathbf{B}_{ij})$, where $\Sigma_{ij}\mathbf{A}_{ij}=\Sigma_{ij}\mathbf{B}_{ij}=1$.

The corresponding objective is to minimize:

$$\begin{aligned} & J(\mathbf{P}, \mathbf{\Lambda}, \mathbf{Z}^{(F)}) = D(\mathbf{W}^{(F)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(F)}) \\ & s.t.\ \mathbf{P} \in \Re_+^{|P|\times K}, \mathbf{\Lambda} \in \Re_+^{K\times K}, \mathbf{Z}^{(F)} \in \Re_+^{K\times D}, \\ & \quad \sum_i \mathbf{P}_{ik} = 1\ \forall k, \sum_k \mathbf{\Lambda}_k = 1 \end{aligned} \quad <2>$$

where $D(\cdot||\cdot)$ is the KL divergence defined above. The constraint that columns of $\mathbf{P}$ must sum to one is added to avoid scaling solutions (e.g., if $\mathbf{P}$ is a solution, $\alpha\mathbf{P}$ can also be the solution if $\mathbf{Z}^{(F)}$ is scaled correspondingly). With the non-negative constraints, this optimization problem is a case of non-negative matrix factorization (NMF) [Lee and Seung 2001].

Users and Tags. Two photos might belong to the same cluster not only due to visual similarity, but also perhaps because they are posted by the same user, or associated with the same tags. For a set of users U, suppose we have a $K\times|U|$ user coefficient matrix $\mathbf{Z}^{(U)}$, where each entry z_{kj} indicates how likely the j-th user posts a photo that falls in the k-th cluster. Similar to the feature matrix factorization, we approximate the photo-user matrix $\mathbf{W}^{(U)}$ by $\mathbf{P}$, $\mathbf{\Lambda}$ and $\mathbf{Z}^{(U)}$ by the following objective:

$$J(\mathbf{P}, \mathbf{\Lambda}, \mathbf{Z}^{(U)}) = D(\mathbf{W}^{(U)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(U)}) \quad <3>$$

subject to $\mathbf{Z}^{(U)} \in \Re_+^{K\times|U|}$, with other constraints and $D(\cdot||\cdot)$ defined as in eq. <2>. Similarly, let $\mathbf{Z}^{(Q)}$ be the tag coefficient matrix, we approximate the photo-tag matrix $\mathbf{W}^{(Q)}$ by $\mathbf{P}$, $\mathbf{\Lambda}$ and $\mathbf{Z}^{(Q)}$ by:

$$J(\mathbf{P}, \mathbf{\Lambda}, \mathbf{Z}^{(Q)}) = D(\mathbf{W}^{(Q)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(Q)}) \quad <4>$$

subject to $\mathbf{Z}^{(Q)} \in \Re_+^{K\times|Q|}$, with other constraints defined as in eq. <2>.

Temporal information. To extract clusters having similar temporal trends, we consider two photos to be in the same cluster if they are posted during the same time slot. For S time slots, let $\mathbf{Z}^{(T)}$ be the time coefficient matrix, where each entry z_{kj} indicates how likely a photo posted at time j belongs to the k-th cluster. We approximate the photo-time matrix $\mathbf{W}^{(T)}$ by $\mathbf{P}$, $\mathbf{\Lambda}$ and $\mathbf{Z}^{(T)}$ as:

$$J(\mathbf{P}, \mathbf{\Lambda}, \mathbf{Z}^{(T)}) = D(\mathbf{W}^{(T)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(T)}) \quad <5>$$

subject to $\mathbf{Z}^{(T)} \in \Re_+^{K\times S}$, with other constraints defined as in eq. <2>. Note that the photo-time matrix contains smoothed temporal information as discussed in section 3. By considering temporal information as one type of relation in the photo streams, it can be dealt with in the same way as in other relations.

Joint objective. Putting together all objective functions with respect to different facets, our objective is to minimize the following function:

$$\begin{aligned} & J(\mathbf{P}, \mathbf{\Lambda}, \{\mathbf{Z}^{(r)}\}) = \sum_{r\in\{F,U,Q,T\}} D(\mathbf{W}^{(r)} \,||\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)}) \\ & s.t.\ \mathbf{P} \in \Re_+^{|P|\times K}, \mathbf{\Lambda} \in \Re_+^{K\times K}, \mathbf{Z}^{(r)} \in \Re_+^{K\times I_r}, \\ & \quad \sum_i \mathbf{P}_{ik} = 1\ \forall k, \sum_k \mathbf{\Lambda}_k = 1 \end{aligned} \quad <6>$$

where $\{\mathbf{Z}^{(r)}\}$ is a set of coefficient matrices for different facets (visual features, users, tags, and times) and I_k denotes the dimensionality of the second dimension of the coefficient matrices. The joint factorization is illustrated in Figure 3(b). Note, eq. <6> can be easily

extended to incorporate additional aspects or to incorporate weights on facets, e.g. a relation that relates a photo by whoever marks it a "favorite" can be simply added to our framework. It is also easy to incorporate other social media contexts, such as the EXIF metadata of images (e.g. image taken time, location and camera settings) whenever they are available. We provide a solution to the joint objective function in the next section.

4.2 Algorithm

We provide a solution to the objective defined in eq. <6>. Since eq. <6> is not convex in all variables, it is difficult to guarantee a global minima solution. By employing the concavity of the log function given in the KL-divergence, we derive a local minima solution to eq. <6> as follows.

Theorem 1. The following update rules will monotonically decrease the cost defined in eq. <6> and therefore converge to an (local) optimal solution to our relational clustering problem:

$$\mathbf{Z}^{(r)}_{kj} \leftarrow \sum_r \sum_i \mathbf{W}^{(r)}_{ij} \mu^{(r)}_{ijk},$$

$$\mathbf{P}_{ik} \leftarrow \sum_r \sum_j \mathbf{W}^{(r)}_{ij} \mu^{(r)}_{ijk},$$

then normalize such that $\sum_i \mathbf{P}_{ik} = 1 \ \forall k$,

$$\mathbf{\Lambda}_k \leftarrow \sum_r \sum_{ij} \mathbf{W}^{(r)}_{ij} \mu^{(r)}_{ijk}, \qquad <7>$$

then normalize such that $\sum_k \mathbf{\Lambda}_k = 1$

where $\mu^{(r)}_{ijk} = \dfrac{\mathbf{P}_{ik}\mathbf{\Lambda}_k\mathbf{Z}^{(r)}_{kj}}{(\mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)})_{ij}}$

The proof of Theorem 1 is provided in the appendix. This iterative update algorithm is a multi-matrices factorization that generalizes the algorithm proposed by Lee and Seung [2001] for solving a single non-negative matrix factorization problem. Table 1 summarizes the process for solving $\mathbf{P}$, $\mathbf{\Lambda}$ and $\{\mathbf{Z}^{(r)}\}$.

Table 1: Algorithm for relational clustering extraction

Algorithm: Relational Soft Clustering
Input: data matrices $\{\mathbf{W}^{(r)}\}$ for $r \in \{F,U,Q,T\}$
Output: $\mathbf{P}$, $\mathbf{\Lambda}$ and $\{\mathbf{Z}^{(r)}\}$
Method:
Initialize $\mathbf{P}$, $\mathbf{\Lambda}$, $\{\mathbf{Z}^{(r)}\}$
Repeat until *convergence*
For each r, update $\mathbf{Z}^{(r)}$ by eq. <7>
update $\mathbf{P}$ and $\mathbf{\Lambda}$ by eq. <7>

Interpretation. We determine the relational clusters based on the solution matrices $\mathbf{P}$, $\mathbf{\Lambda}$ and $\{\mathbf{Z}^{(r)}\}$. Specifically, the soft membership (ref. section 4.1) of each photo i in community k is determined by the conditional probability $P(k|i) = P(i,k)/P(i)$, where $P(i,k)$ is given by $(\mathbf{P}\mathbf{\Lambda})_{ik}$, and the marginal probability $P(i)$ is given by $\Sigma_k(\mathbf{P}\mathbf{\Lambda})_{ik}$. The soft membership of a user or a tag can be computed in the same way with corresponding normalized coefficient matrices. Such soft membership not only provides information about the relationship between an object and a cluster; it can also be used to infer the relationship between two entities or two clusters in the relational structure. For example, the joint probability between a photo i and a tag j is computed by $P(i,j) = \Sigma_k \mathbf{\Lambda}_k\mathbf{P}_{ik}\mathbf{Z}_{kj}{}^{(Q)}$. The joint probability of two clusters k and l can be determined by the marginal probability $P(k,l) = \Sigma_i P(k|i)P(l|i)P(i)$, where i is the entity index within a particular facet such as photo or tag facet. In some application, e.g. visualization, where hard membership (disjoint clusters) may be used, we can convert soft membership to hard membership by choosing the maximum $P(k|i)$ over k.

Computational complexity. We now investigate the time complexity for each iteration of the updates in eq. <7>. The most time-consuming part is to compute $(\mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)})_{ij}$ $\forall i,j,r$. However, most data matrices $\mathbf{W}^{(\cdot)}$ are sparse, i.e. they have few non-zero entries. Hence we only need to compute the corresponding $(\mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)})_{ij}$ for each non-zero entry (i,j) in $\mathbf{W}^{(r)}$, which takes O(K) time with K being the number of clusters. Let n denote the largest number of non-zero entries of all data matrices, the total time complexity is O(nK). If we consider K is bounded by some constants, the time complexity per iteration is linear in O(n), the number of non-zero entries in data matrices. Note that the sparse property might not hold for the photo-feature matrix $\mathbf{W}^{(F)}$; however, $\mathbf{W}^{(F)}$ is typically constructed based on fixed length feature vectors. The number of non-zero entries in $\mathbf{W}^{(F)}$ only depends on the number of photos and hence it can be considered to have the same degree of sparseness as other data matrices.

5. RELATIONAL MODULARITY

We discuss how to determine the number of relational clusters in this section. So far we have assumed the number of relational clusters, K, is given. However, since it is almost impossible to always know the number of clusters in a large network beforehand, such an assumption will limit the scope of application of our framework. In the following we propose an automatic mechanism to determine the number of relational clusters.

Let us first examine the relationship between the number of clusters and the clustering objective defined in eq. <6>. The objective is defined based on KL divergence and there exists a relationship between mutual information and the KL divergence. We can show that the KL divergence between the observed joint distribution and the predicted (i.e. estimated by certain model) joint distribution can be expressed in terms of loss in mutual information, i.e.,

$$I(X;Y) - I(\hat{X};\hat{Y}) = D(p(X,Y) \,\|\, q(X,Y)) \qquad <8>$$

where the $(\hat{X}, \hat{Y})$ is a mapping from (X, Y) via the soft-clustering algorithm. There is a straightforward proof to extend the results in [Dhillon et al. 2003]. $p(X,Y)$ and $q(X,Y)$ are the observed and predicted joint distribution of X and Y, respectively. $I(X;Y)$ denotes the mutual information between X and Y. In our case, the observed joint distribution is given by a data matrix $\mathbf{W}^{(r)}$, the predicted joint distribution is given by the factorization $\mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)}$, and $(\hat{X},\hat{Y}) = \{(\hat{x}_1,\hat{y}_1),(\hat{x}_2,\hat{y}_2),...(\hat{x}_K,\hat{y}_K)\}$ represents K disjoint clusters where $\hat{x}$ and $\hat{y}$ contain a subset of X and Y respectively. Based on this mutual information interpretation, our multi-relational clustering objective can be viewed as finding a clustering structure such that the information about multiple relations among facets remains as much as possible in the structure. Since $I(X;Y)$ is constant given the data, minimizing the KL divergence is equivalent to maximizing the mutual information $I(\hat{X};\hat{Y})$ subject to the number of clusters. The clustering objective based on the KL divergence is not sufficient to determine the optimal number of clusters, since it doesn't include the model penalty. To find a structure with only a small set of clusters, we introduce a penalized term, the mutual information conditioned on the clustering:

$$I(X;Y \mid C) = \sum_{c\in C}\sum_{x\in X}\sum_{y\in Y} p(x,y,c)\log\frac{p(c)\,p(x,y,c)}{p(x,c)\,p(y,c)}, \qquad <9>$$

where X and Y are two set of objects, $C=\{1, 2, \ldots, K\}$ is the set of cluster indices.

To evaluate the quality or goodness of a relational structure, we define a mutual information based function called *relational modularity* Q_I, as:

$$\begin{aligned} Q_I(K) &= \sum\nolimits_r I(X^{(r)};Y^{(r)} \mid S_K) - D(\mathbf{W}^{(r)} \,\|\, \mathbf{P}\mathbf{\Lambda}\mathbf{Z}^{(r)}) \\ &= (\sum\nolimits_r I(X^{(r)};Y^{(r)} \mid S_K)) - J(\mathbf{P},\mathbf{\Lambda},\{\mathbf{Z}^{(r)}\}) \end{aligned} \qquad <10>$$

where $X^{(r)}$ and $Y^{(r)}$ are two corresponding facets of the relation r, represented by matrix $\mathbf{W}^{(r)}$, J is the clustering objective function defined in e.q. <6> subject to number of cluster K, S_K is the K-clustering structure obtained from solving J. The computation of $Q_I(K)$ involves two steps: First, for a given K, solve J by the algorithm in Table 1, and then use the solution matrices ($\mathbf{P}$, $\mathbf{\Lambda}$, $\{\mathbf{Z}^{(r)}\}$) to compute the conditional mutual information $I(X^{(r)}; Y^{(r)}|S_K)$ based on e.q. <9>, where the joint probabilities $p(x,c)$, $p(y,c)$ and $p(x,y,c)$ are given by:

$$p(x,c) \propto P(c \mid x) \sum_{y} \mathbf{W}_{xy}^{(r)},$$
$$p(y,c) \propto P(c \mid y) \sum_{x} \mathbf{W}_{xy}^{(r)}, \qquad <11>$$
$$p(x,y,c) \propto \mathbf{W}_{xy}^{(r)} P(c \mid x) P(c \mid y),$$

where $P(c|x)$, $P(c|y)$ can be obtained from the solution matrices as described in section 4. The best clustering structure is given by the maximal Q_I.

In the following discussion we connect the relational modularity to the modularity introduced in [Newman and Girvan 2004]. They define modularity Q as:

$$Q(K) = \sum_{k=1}^{K} l_{kk} / L - \left(l_k / 2L \right)^2 \qquad <12>$$

where K is the number of clusters, L is the total number of links in the network, l_{kk} is the number of links between nodes belonging to cluster k, l_k is the total degree of nodes in cluster k (i.e. number of links with at least one end falls in cluster k). The idea is to divide the network such that the number of links within clusters is higher than expected, and hence the modularity Q measures the deviation between fraction of edges within communities (expressed by the first term inside the summation) and the expected fraction of such edges (expressed by the second term). It can be seen that the first term corresponds to the joint probability $p(x,y,c)$, and the second term corresponds to the marginal probabilities $p(x,c)$ and $p(y,c)$. Hence our relational modularity naturally extends the modularity concept by using the conditional mutual information.

6. EXPERIMENTS

This section reports our experimental studies on a Flickr group dataset. We first describe the dataset used in our experiments (section 6.1), and present case studies on the relational clusters extracted from the data (section 6.2). Finally we quantitatively evaluate and discuss the quality of the clustering structure through a prediction task (section 6.3). In the appendix, we evaluate the effectiveness of relational modularity by using synthetic datasets, and we present a qualitative user study to evaluate the benefits of our framework in exploring group photo streams. Figure 4 presents a case study for which clustering results are recognized by the participants in our user study.

6.1 Flickr Dataset

Using the Flickr API[2], we collect data from a sample of 120 Flickr groups based on the group size distribution. We download all publicly available photos for each group. Our dataset consists of 111,108 photos, 8,117 unique users, and 102,607 unique tags in total. The photo post times range from January 1, 2004 to January 8, 2009, enabling us to analyze long-term temporal patterns in this collection.

[2] http://www.flickr.com/services/api/

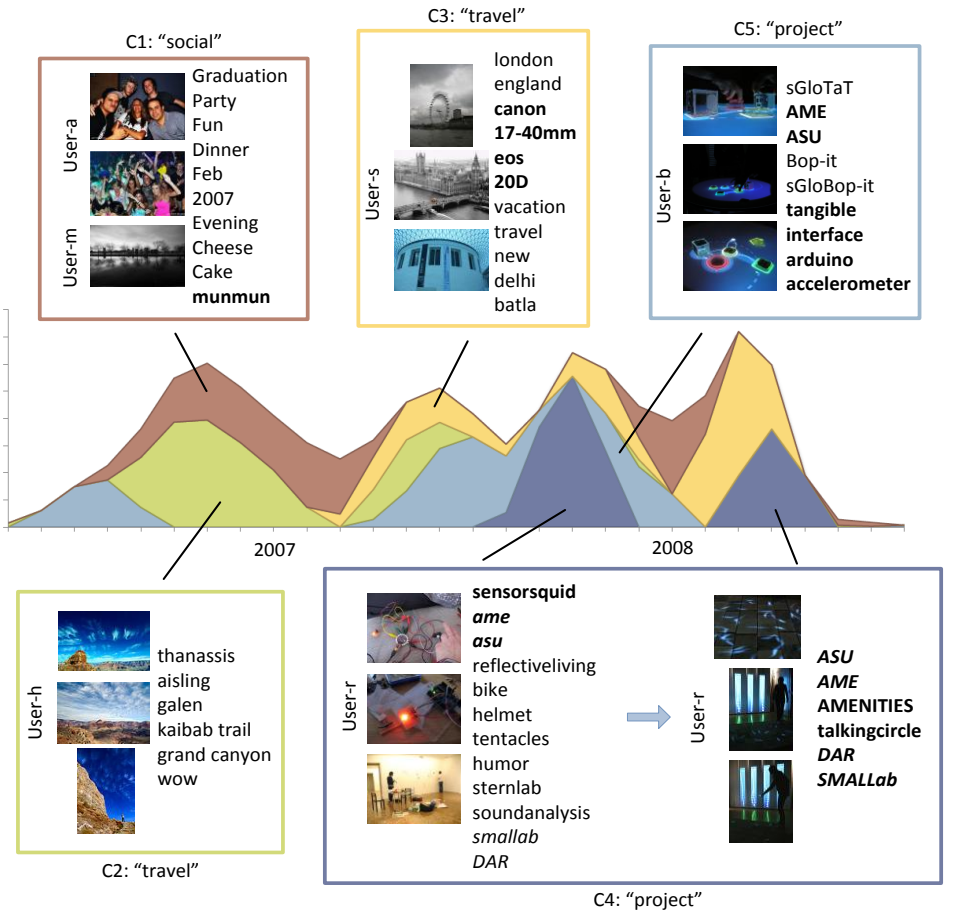

Figure 4: This figure presents a case which results are recognized by the participants in our user study (see appendix). The 5-cluster results extracted from the "AME" group. We display representative photos, users (renamed for privacy consideration) and tags, as well as the temporal strength of each cluster over time. Based on participants' feedback, we annotate the clusters as C1: "social," C2: "travel," C3: "travel," C4: "project" and C5: "project." Based on the photos uploaded by the group members, the clustering results capture interesting evolution of the activity in the local community. For example, social events (as in C1) were more frequently in 2007. Travel photos (as in C2 and C3) have been continuously uploaded by different users. Users started documenting their work in Fall 2007 (as in C4 and C5). C4 captures a fine-grained activity evolution – it represents that the project work involving user-r in different points of time are co-located (e.g. "smallab" and "DAR" in C4 are location names), but the project content are different.

For comparison, we randomly select 1000, 2000, 5000 and 10,000 photos from the dataset and create 4 random groups.

6.2 Clustering Results

We investigate the clustering structures extracted from those groups and present the results for four groups: (a) "35mm Focal Length (** NOT 35mm film... **)[3]" (4024 photos), (b) "The Southwest United States[4]" (4968 photos), (c) "Laser photography[5]" (1023 photos) and (d) the largest random group which consists of 10,000 randomly selected photos. For brevity we denote these groups as group "A", "B", "C" and "D".

[3] http://www.flickr.com/groups/27044956@N00
[4] http://www.flickr.com/groups/10477049@N00
[5] http://www.flickr.com/groups/31293421@N00

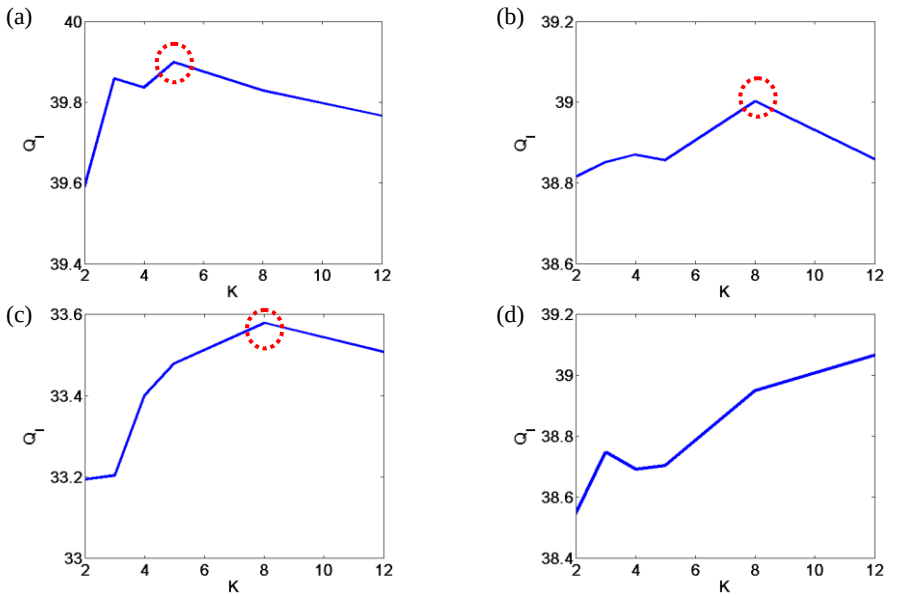

Figure 5: The best clustering structure for each group is determined based on the maximal relational modularity Q_I. (a)—(d) plot the values of Q_I over the number of clusters for group A, B, C and D.

We focus on the following questions:

(1) How many clusters do we need for best capturing the relational structure?
(2) Can we extract relational structures across different facets?
(3) What are representative photos or tags, in the group?
(4) What is the temporal aspect of a group structure, i.e., how likely does a cluster appear during a particular time period?
(5) What are the relationships among clusters?

We first determine the best clustering structure for each group by plotting the relational modularity Q_I over the number of clusters K. The effectiveness of Q_I is studied in the appendix. The result for the group A is shown in Figure 5(a). As can be seen, within a certain range of K (2, …, 12), Q_I has the highest value when K=5, which is considered to be the best clustering structure. Figure 6 illustrates the clustering results of group A which comprises 5 clusters. We show the most representative photos for a cluster (ref. Figure 6(ii)) based on how likely a photo i belongs to the cluster k, i.e. p_{ik} in **P**. The temporal strength of the clustering structure, i.e., how likely a cluster appears at a particular time (ref. Figure 6(i)), is determined based on the coefficient matrix $\mathbf{Z}^{(T)}$. As can be seen, some clusters tend to be bursty at certain time periods, e.g., A2 and A3, while some clusters are more stable over time, e.g., A4. This might be explained by the tag aspect. In Figure 6(iv), we list the top tags of each cluster based on the values in the coefficient matrix $\mathbf{Z}^{(Q)}$. From the tag list, we observe that stable clusters like A4 tend to have tags signifying photographic style (e.g., "blackandwhite") or camera brand (e.g., "konica"), as opposed to tags signifying the spatial context of the photos (e.g. "brasil" in A1, "zeiss" in A3, "street" in A4, etc.). Note that the top users of each cluster can be determined based on $\mathbf{Z}^{(U)}$, but here we omit the list of their identifiable names. To see if the clustering structure is consistent across different facets, we plot the photo-tag and photo-user matrices where the rows and columns are re-ordered based on the cluster indices of the corresponding photos, tags or users. Here the cluster indices are determined based on the maximal posterior probability computed as described in section 4. The clustering structure is revealed in both matrices (ref. Figure 6(iii)).

We also examine the relationship between clusters by computing the joint probability between any two clusters based on different facets. Figure 7 shows the relationships

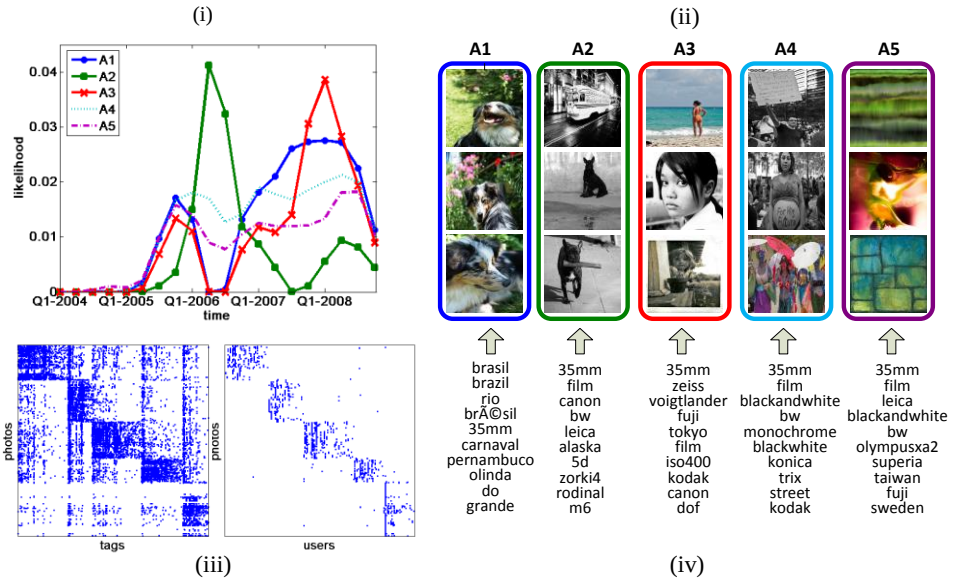

Figure 6: Group A "35mm Focal Length" has 5 clusters. (i) The temporal strength of the clustering structure, i.e., how likely a cluster appears at a particular time, is determined based on the coefficient matrix $\mathbf{Z}^{(T)}$. (ii) The representative photos for each cluster based on how likely a photo i belongs to the cluster k, i.e. p_{ik} in $\mathbf{P}$. (iii) The photo-tag and photo-user matrices where the rows and columns are re-ordered based on the cluster indices of the corresponding photos, tags or users. The clustering structure is revealed in both matrices. (iv) The top tags of each cluster based on the values in the coefficient matrix $\mathbf{Z}^{(Q)}$. The tags reflect photographic, style (e.g., "blackandwhite"), camera brand (e.g., "konica"), or the spatial context of the photos (e.g. "brasil" in A1, "zeiss" in A3, "street" in A4, etc.).

among clusters within group A, where thicker lines indicate higher joint probability of elements in a particular facet. It can be seen that all five clusters have stronger within-cluster joint probability in almost all facets, indicating a strong relational clustering structure in this group.

We apply the same analysis on other groups. Figure 5(b) and (c) show that both groups B and C have the best structures with 8 clusters. The temporal strength of the

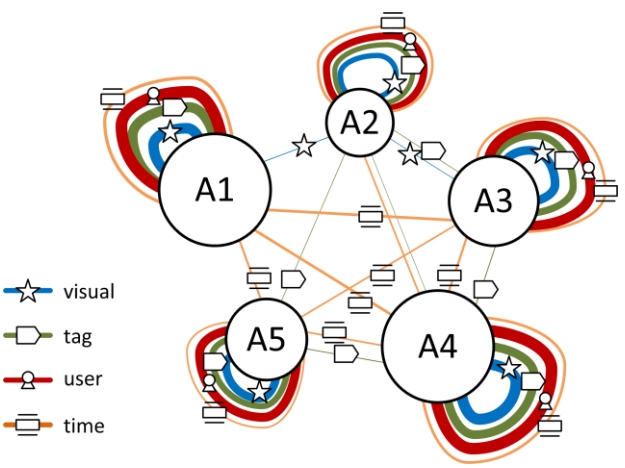

Figure 7: The relationships among clusters within group A. We compute the joint probability between every two clusters based on different relations (visual, tag, etc.) Thicker lines indicate higher joint probability of elements in a particular facet. Only lines with probability larger than a threshold are shown. It can be seen that all five clusters have stronger within-cluster joint probability in most facets, indicating a strong relational clustering structure in this group.

structures and the representative photos for each cluster are also shown. In Figure 8, we observe an active period for group B – during the 2007 year, several clusters are likely to co-exist (ref. Figure 8(i)). Although group C has the same number of clusters, it exhibits different temporal characteristics – most clusters in group C are likely to appear only at a certain time period (ref. Figure 9(i)). We observe similar consistent clustering structures in their re-ordered photo-tag and photo-user matrices.

To understand the limitation of our structure discovery approach, we apply the same analysis on the random groups. The random group D exhibits characteristics that are very different from normal user groups A, B and C. First, given a range of number of clusters K, the relational modularity Q_I tends to increase with K (ref. Figure 5 (d)). In Figure 10, the top photos of each cluster (except for D3) in group D do not give as a clear sense of the cluster as those in other groups (ref. Figure 10(ii)). Similar characteristics have been observed from other random groups.

Figure 8: Group B: "The Southesst United States" comprises 8 clusters. The group has an active period – during the 2007 year several clusters are likely to co-exist. The cluster representative photos tend to have similar scenes.

Figure 9: Group C: "Laser Photography" comprises 8 clusters. It has the same number of clusters with group B but exhibits different temporal characteristics – most clusters in group C are likely to appear only at a certain time period.

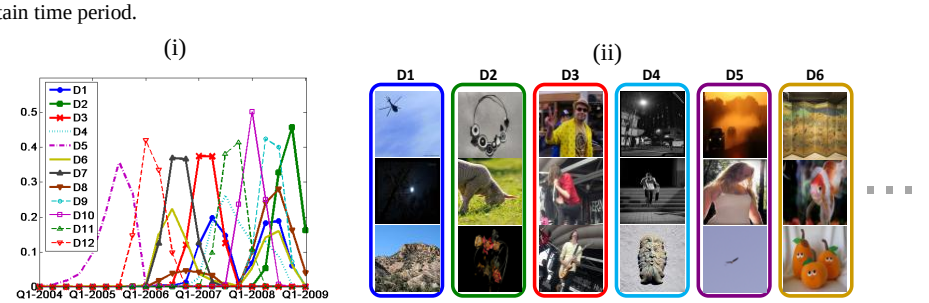

Figure 10: Group D: Random group. The random group D exhibits characteristics that are very different from normal user groups A, B and C. The top photos of each cluster (except for D3) in group D do not give as a clear a sense of the cluster as those in other groups. The photo-tag and photo-user matrices suggest that although our algorithm gives a consistent clustering structure, the structure looks more artificial because of the almost uniform cluster sizes.

The clustering results suggest that our algorithm is able to extract meaningful structures that characterize the relational patterns in group photo streams, in terms of their representative photos, tags and users, as well as the time profile of the clusters. However, it is also possible that the algorithm gives an ad-hoc structure that makes little sense, as in the random group case.

6.3 Evaluation via Prediction

How can we quantify the meaningfulness of an extracted structure? We design a prediction task to examine this question. The idea is if a structure captures recurring relational patterns, it should be able to predict missing relations in the same group photo stream. Hence we design a task to evaluate a relational structure by its predictability, that is, a structure is meaningful if the relations of a photo in the group photo stream can be predicted by the structure.

Prediction setting. We design a task for predicting the relations of unseen photos, given the extracted structure. For each group, we randomly choose 70% and 90% of the photos for extracting the clustering structure (training), and use the remaining 30% and 10% of the photos for testing. The task is to predict the photo-tag relation, i.e. tags associated with the testing images. Our prediction utilizes the coefficient matrices obtained from the training stage, with an estimation of p_{ik} for the test photos using a folding-in technique [Schein et al. 2002]. We determine if a photo p_i will be tagged with a tag x_j by the conditional probability:

$$P(x_j \mid p_i) \propto P(x_j, p_i) \approx \sum_k \mathbf{\Lambda}_k \cdot \mathbf{P}_{ik} \cdot \mathbf{Z}^{(Q)}_{kj}$$

where $\mathbf{\Lambda}$, $\mathbf{P}$ and $\mathbf{Z}$ are defined as in section 4. Our method is denoted by "RSC" (Relational Soft Clustering).

Baseline methods. We compare our prediction results with three baseline methods: (a) feature-based prediction (denoted by "Features") – predicting tags from photos having most similar visual features (i.e. nearest neighbor); (b) tag-based (denoted by "Tags") – predicting tags based on the tag frequency; (c) feature/tag (denoted by "F/T") – predicting tags by only using the feature and tag information in joint factorization.

Evaluation metrics. We use the following metrics adopted in Information Retrieval:

(1) **S@10** (the success among the top 10 results): S@10 is defined as the probability of at least one correct tag among the top 10 results.
(2) **P@10** (the precision of the top 10 results): P@10 is defined as the proportion of predicted tags that is correct, averaged over all photos.
(3) **MRR** (mean reciprocal rank): MRR measures the ability of a method to return a correct tag at the top of the ranking. It is proportional to $\Sigma_i\ 1/r_i$, where r_i is the rank of the first correct tag for the i-th photo.
(4) **NDCG** (Normalized Discount Cumulative Gain [Järvelin and Kekäläinen 2000]): One advantage of the measure is its sensitivity to the prediction order. The NDCG is proportional to $\Sigma_i \delta(i)/\log(1+i)$, where i is the rank of predicted tags, $\delta(i)=1$ if the prediction of the rank-i tag is correct and 0 otherwise. Unlike MRR which only considers the first correct tag, NDCG considers multiple correct tags at the top of the ranking.

Results and discussion. We provide the detailed results in the appendix (ref. Table 2 and Table 3). The result shows that our method significantly outperforms all baselines by 35%–420% (based on NDCG), or 44%–390% (based on P@10) on an average. For comparison, Figure 11 shows the relative improvement of prediction performance – each method is compared against the first baseline (Features) method.

There are several observations:

(1) Prediction based on visual features alone performs the worst.
(2) Prediction based on tag frequency works better, but the performance is poor after the most relevant tag. This can be seen from its high performance in terms of S@10 and

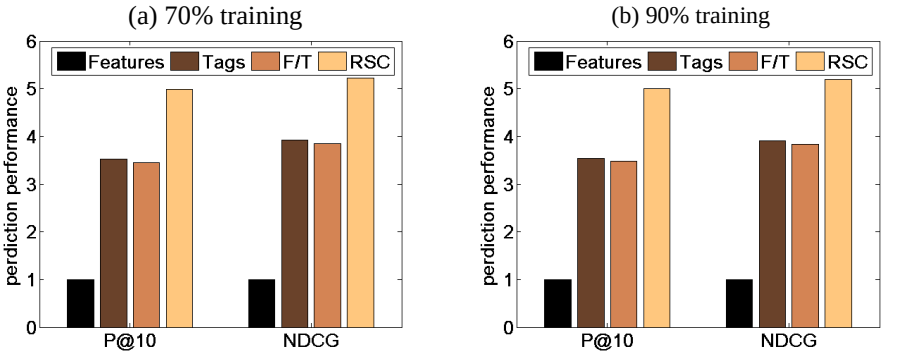

Figure 11: Relative prediction performance for (a) 70% photos for training and (b) 90% photos for training. We compare our prediction method (RSC) with three baseline methods, feature-based, tag based and both (F/T). Our method significantly improves the tag prediction in group photo streams.

MRR, but low performance in terms of P@10 and NDCG. The reason for this is that many photos posted in a group are also associated with one or two group related tags, e.g. "35mm" in group A. Hence, the metric P@10 and NDCG are more effective in differentiating prediction qualities in this task.

(3) Combining feature and tag performs similarly as tag-frequency based prediction.

(4) By incorporating various relations that consist of visual and contextual information (photo-tag, photo-user, photo-time, and photo-feature), our joint structure analysis finds the highest quality tags for the testing photos. This may be attributed to the "event locality" in Flickr photos, i.e., many photos are well correlated to either global events, or to events that are observed by the users. This implies that the use of tags is highly correlated to the event context which is sensitive to a particular user, time, visual appearance, etc. The relational data model serves as a way to capture the event context.

The results suggest that our analysis based on relational clustering structure help improve the quality of tag prediction. The experiment thus provides a quantifiable sense about the meaningfulness of the extracted structure.

7. OPEN ISSUES / EXTENSIONS

The proposed framework has several open issues and can lead to potential extensions:

- In this work, we impose a fixed form of structure, i.e., clustering structure, to characterize a group photo stream. However, there might exist other forms that better capture the data, including generative forms from which the specific structure observed in the data is derived, e.g. chain, multi-level hierarchy, etc., as suggested in [Kemp and Tenenbaum 2008].
- In our factorization method, we use the product of the facet matrices to fit each observed relation. A more natural extension is to use kernel representation for those factors to exploit their non-linear relationship.
- The proposed joint factorization method can be easily expanded to add other features, which allows this method to be easily incorporated other social media contexts, such as the EXIF metadata of an image (e.g. image taken time, location and camera settings), or the content filtering labels (e.g. "politics" or "business") associated with a blog post, etc., whenever they are available.
- With richer contextual information such as image taken time and location, the proposed framework can be applied to event-centric multimedia application, such as detecting events (e.g. attending a performance) based on the uploaded photos (and the associated contexts) and automatically recommending photos relevant to the events (e.g. photos about the performers or similar performances).

- Moving beyond event-centric application, another interesting direction is community-centric application. Our proposed framework can be used to detect so called "boundary objects" which are "weakly structured in common use, and become structured in individual-site use" [Star and Griesemer 1989]. The concept is illustrated in Figure 7, where the relational co-occurrences are evident within clusters but faint between clusters. The detection of boundary objects can be useful to support and sustain the sharing practices in a social media platform. Such system could provide different navigation mechanisms for different types of sharing space, e.g. color-based navigation function for exploring photos about a common subject such as "sky," or time-based navigation function for exploring photos about people and events within local communities.

8. CONCLUSION

We presented a method for extracting multi-relational structure in social media streams. Structure discovery is a fundamental problem with applications in content organization, recommendation systems and exploratory social network analysis. We used a nonnegative joint matrix factorization approach to find a set of soft clusters that reflect various relations in a group photo stream. By leveraging various relations, our method dealt with visual and contextual information, including visual content features, photo tags, owners and post times, in a unified manner.

We provided an efficient algorithm to solve the clustering problem that scales linearly with the data size. The discovered structured are interpretable in terms of minimizing the mutual information of the joint distribution. A relational modularity function was proposed to determine model penalty and estimate the optimal number of clusters. Extensive experiments on a Flickr dataset and user study show that (a) our analysis can capture the dynamics of group patterns, and give meaningful summary of group photo streams; (b) compared with baseline methods, our joint analysis performs the highest quality tag prediction. These results indicate the utility of our relational clustering method.

As part of our future work, we plan to extend the current framework to (a) consider different forms of the structures (b) and extend the current matrix factorization framework to tensor based analysis.

ACKNOWLEDGEMENTS

We thank many artists who have shared their photos under a Creative Commons license. This article includes photos from a number of sources attributed at: http://bit.ly/vWA00P

REFERENCES

AHERN, S., NAAMAN, M., NAIR, R. AND YANG, J. 2007. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. *Proc. of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 10.

BACKSTROM, L., HUTTENLOCHER, D., KLEINBERG, J. AND LAN, X. 2006. Group formation in large social networks: Membership, growth, and evolution. *SIGKDD*, 2006, ACM Press, 44-54.

BANERJEE, A., BASU, S. AND MERUGU, S. 2007. Multi-way clustering on relation graphs. *SDM*, 2007.

BEKKERMAN, R., EL-YANIV, R. AND MCCALLUM, A. 2005. Multi-way distributional clustering via pairwise interactions. *ACM Intl. Conf. Proc. Series*, 41-48.

BLEI, D., NG, A. AND JORDAN, M. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research 3*: 993-1022.

BLEI, D. AND LAFFERTY, J. 2006. Dynamic topic models. *ICML*, ACM, 120.

CAI, D., HE, X., LI, Z., MA, W. AND WEN, J. 2004. Hierarchical clustering of www image search results using visual, textual and link information. *ACM MM*, 2004, ACM New York, NY, USA, 952-959.

CHEN, H., CHANG, M., CHANG, P., TIEN, M., HSU, W. AND WU, J. 2008. Sheepdog: Group and tag recommendation for flickr photos by automatic search-based learning.

DHILLON, I., MALLELA, S. AND MODHA, D. 2003. Information-theoretic co-clustering. *SIGKDD*, ACM New York, NY, USA, 89-98.

DOREIAN, P. AND FUJIMOTO, K. 2001. Structures of supreme court voting. *University of Pittsburgh, manuscript, version November 3*: 2001.
GARG, N. AND WEBER, I. 2008. Personalized, interactive tag recommendation for flickr. *RecSys*, 2008, ACM New York, NY, USA, 67-74.
JÄRVELIN, K. AND KEKÄLÄINEN, J. 2000. Ir evaluation methods for retrieving highly relevant documents. *SIGIR*, 2000, ACM New York, NY, USA, 41-48.
KEMP, C. AND TENENBAUM, J. 2008. The discovery of structural form. *PNAS 105* (31): 10687.
KENNEDY, L., NAAMAN, M., AHERN, S., NAIR, R. AND RATTENBURY, T. 2007. How flickr helps us make sense of the world: Context and content in community-contributed media collections. *ACM MM*, 2007, ACM New York, NY, USA, 631-640.
KIRSTEN, M. AND WROBEL, S. 1998. Relational distance-based clustering. *Inductive logic programming: ILP-98, Madison, Wisconsin, USA, July 22-24, 1998: proceedings*, Springer Verlag, 261.
KUMAR, R., NOVAK, J. AND TOMKINS, A. 2006. Structure and evolution of online social networks. *SIGKDD*, Philadelphia, PA, USA, 2006, ACM Press, 611-617.
LEE, D. AND SEUNG, H. 2001. Algorithms for non-negative matrix factorization. *NIPS*, 2001, 556–562.
LI, T. AND ANAND, S. 2007. Diva: A variance-based clustering approach for multi-type relational data. *SIGKDD*, 2007, ACM New York, NY, USA, 147-156.
LIN, Y.-R., CHI, Y., ZHU, S., SUNDARAM, H. AND TSENG, B.L. 2008. Facenet: A framework for analyzing communities and their evolutions in dynamics networks. *WWW*, 2008, ACM Press.
LIN, Y.-R., SUNDARAM, H., DE CHOUDHURY, M. AND KELLIHER, A. 2009a. Temporal patterns in social media streams: Theme discovery and evolution using joint analysis of content and context. *IEEE ICME*, 2009.
LIN, Y.-R., SUNDARAM, H. AND KELLIHER, A. 2009b. JAM: Joint action matrix factorization for summarizing a temporal heterogeneous social network. *ICWSM*, 2009.
LIU, Z. AND LAGANIERE, R. 2007. Phase congruence measurement for image similarity assessment. *Pattern Recogn. Lett. 28* (1): 166-172.
LONG, B., WU, X., ZHANG, Z. AND YU, P. 2006. Unsupervised learning on k-partite graphs. *SIGKDD*, 2006, ACM Press New York, NY, USA, 317-326.
LOWE, D.G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60* (2): 91-110.
LOY, G. AND ZELINSKY, A. 2003. Fast radial symmetry for detecting points of interest. *IEEE Trans. Pattern Anal. Mach. Intell. 25* (8): 959-973.
MCCOWAN, L., GATICA-PEREZ, D., BENGIO, S., LATHOUD, G., BARNARD, M. AND ZHANG, D. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Trans. on Pattern Analysis and Machine Intelligence 27* (3): 305-317.
NEGOESCU, R. AND GATICA-PEREZ, D. 2008. Analyzing flickr groups. *CIVR*, 2008, ACM New York, NY, USA, 417-426.
NEWMAN, M. AND GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Physical Review E 69* (2): 26113.
PALLA, G., BARABASI, A. AND VICSEK, T. 2007. Quantifying social group evolution. *eprint arXiv: 0704.0744*.
REGE, M., DONG, M. AND HUA, J. 2008. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering.
SCHEIN, A., POPESCUL, A., UNGAR, L. AND PENNOCK, D. 2002. Methods and metrics for cold-start recommendations. *SIGIR*, ACM New York, NY, USA, 253-260.
SHAMMA, D., SHAW, R., SHAFTON, P. AND LIU, Y. 2007. Watch what i watch. *MIR*, 2007, ACM.
SIGURBJÖRNSSON, B. AND VAN ZWOL, R. 2008. Flickr tag recommendation based on collective knowledge.
STAR, S. AND GRIESEMER, J. 1989. Institutional ecology,'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39. *Social studies of science 19* (3): 387-420.
SUN, J., FALOUTSOS, C., PAPADIMITRIOU, S. AND YU, P. 2007. Graphscope: Parameter-free mining of large time-evolving graphs. *SIGKDD*, 2007, ACM Press, 687-696.
TANG, L., LIU, H., ZHANG, J. AND NAZERI, Z. 2008. Community evolution in dynamic multi-mode networks. *SIGKDD*, 2008, ACM New York, NY, USA.
TONG, H., HE, J., LI, M., ZHANG, C. AND MA, W. 2005. Graph based multi-modality learning. *ACM MM*, ACM New York, NY, USA, 862-871.
WANG, X. AND MCCALLUM, A. 2006. Topics over time: A non-markov continuous-time model of topical trends. *SIGKDD*, ACM, 433.
WANG, X., SUN, J., CHEN, Z. AND ZHAI, C. 2006. Latent semantic analysis for multiple-type interrelated data objects. *SIGIR*, 2006, ACM Press, 236-243.
XIAO, Z., HOU, Z., MIAO, C. AND WANG, J. 2005. Using phase information for symmetry detection. *Pattern Recogn. Lett. 26* (13): 1985-1994.
XIE, L., CHANG, S., DIVAKARAN, A. AND SUN, H. 2002. Structure analysis of soccer video with hidden markov models. *ICASSP*, IEEE; 1999, 4096-4099.
ZHU, S., YU, K., CHI, Y. AND GONG, Y. 2007. Combining content and link for classification using matrix factorization. *SIGIR*, 2007, ACM Press, 487-494.
ZUNJARWARD, A., SUNDARAM, H. AND XIE, L. 2007. Contextual wisdom: Social relations and correlations for multimedia event annotation. *ACM MM*, 2007.

APPENDIX A. Proof of Theorem 1

We employ the concavity of log function to prove the correctness of eq. <7>. Because $\log(\sum_k a_{ik}b_{kj})$ is a convex function, the following equality holds for all i, j, and $\sum_k v_{ijk} = 1$:

$$-\log\left(\sum_k a_{ik}b_{kj}\right) \le -\left(\sum_k v_{ijk} \log \frac{a_{ik}b_{kj}}{v_{ijk}}\right), \text{ where } v_{ijk} = \frac{a_{ik}b_{kj}}{\sum_k a_{ik}b_{kj}}$$

Hence, we have:

$$\begin{aligned} J(\mathbf{P},\mathbf{\Lambda},\{\mathbf{Z}^{(r)}\}) &= \sum_r \sum_{ij} \left(-\mathbf{W}_{ij}^{(r)} \log \sum_k \mathbf{P}_{ik}\mathbf{\Lambda}_k \mathbf{Z}_{kj}^{(r)} + \sum_k \mathbf{P}_{ik}\mathbf{\Lambda}_k \mathbf{Z}_{kj}^{(r)} \right) + const \\ &\le \sum_r \sum_{ijk} -\mathbf{W}_{ij}^{(r)} \mu_{ijk}^{(r)} \log \frac{\mathbf{P}_{ik}\mathbf{\Lambda}_k \mathbf{Z}_{kj}^{(r)}}{\mu_{ijk}^{(r)}} + \mathbf{P}_{ik}\mathbf{\Lambda}_k \mathbf{Z}_{kj}^{(r)} + const \\ &\stackrel{def}{=} \mathcal{Q}(\mathbf{P},\mathbf{\Lambda},\{\mathbf{Z}^{(r)}\};\{\mu_{ijk}^{(r)}\}) \end{aligned}$$

where $\mu^{(r)}$ is defined as in eq. <7>. With the constraints $\sum_i \mathbf{P}_{ik} = 1$, the Lagrangian of $\mathcal{Q}$ is defined as:

$$L = \mathcal{Q}(\mathbf{P},\mathbf{\Lambda},\{\mathbf{Z}^{(r)}\};\{\mu_{ijk}^{(r)}\}) + \varepsilon_{\mathbf{P}}\left(\sum_i \mathbf{P}_{ik} - 1\right) + \varepsilon_{\mathbf{\Lambda}}\left(\sum_k \mathbf{\Lambda}_k - 1\right)$$

Update $\mathbf{Z}^{(r)}$: with $\mathbf{P}$ and $\mathbf{\Lambda}$ fixed, we have:

$$\frac{\partial L}{\partial \mathbf{Z}_{kj}^{(r)}} = \sum_r \sum_i -\mathbf{W}_{ij}^{(r)} \mu_{ij}^{(r)} / \mathbf{Z}_{kj}^{(r)} + const = 0$$

By solving this equation, we obtain the update rule for $\mathbf{Z}^{(r)}$.
Update $\mathbf{P}$: with $\{\mathbf{Z}^{(r)}\}$ and $\mathbf{\Lambda}$ fixed, we have:

$$\frac{\partial L}{\partial \mathbf{P}_{ik}} = \sum_r \sum_j -\mathbf{W}_{ij}^{(r)} \mu_{ij}^{(r)} / \mathbf{P}_{ik} + \varepsilon_{\mathbf{P}} + const = 0$$

$$\frac{\partial L}{\partial \varepsilon_{\mathbf{P}}} = \sum_i \mathbf{P}_{ik} - 1 = 0$$

By solving the equations, we obtain the update rule for $\mathbf{P}$.
Update $\mathbf{\Lambda}$: with $\{\mathbf{Z}^{(r)}\}$ and $\mathbf{P}$ fixed, we have:

$$\frac{\partial L}{\partial \mathbf{\Lambda}_k} = \sum_r \sum_{ij} -\mathbf{W}_{ij}^{(r)} \mu_{ij}^{(r)} / \mathbf{\Lambda}_k + \varepsilon_{\mathbf{\Lambda}} + const = 0$$

$$\frac{\partial L}{\partial \varepsilon_{\mathbf{\Lambda}}} = \sum_i \mathbf{\Lambda}_k - 1 = 0$$

By solving the equations, we obtain the update rule for $\mathbf{\Lambda}$. □

APPENDIX B. Effectiveness of Relational Modularity

To study the effectiveness of relational modularity Q_I, we generate a family of tripartite networks by the following process. Each dataset contains N entities in each of three different dimensions. These entities belong to C clusters. Each entity i of facet 1 will attach m edges to facet 2 and m edges to facet 3. The entities being attached are selected as follows: with probability 1-p_{out}, entity i will randomly connect to an entity within the same cluster, and with probability p_{out}, entity i will randomly connect to an entity outside its cluster. The out-linking probability p_{out} can be viewed as the degree of noise to an ideal clustering structure. The simulation is similar to the one described in [Newman and Girvan 2004], except for we extend the idea to a multi-relational network setting.
We experiment on the synthetic networks generated with number of entities N=128 for each facet, and different parameter values for $p_{out} \in \{0.1, 0.2, 0.3\}$ and $C \in [2,8]$. We fix m to be half of N.

We compare the effectiveness of Q_I with the modularity function (denoted as Q) proposed in [Newman and Girvan 2004]. The effectiveness is evaluated based on the true

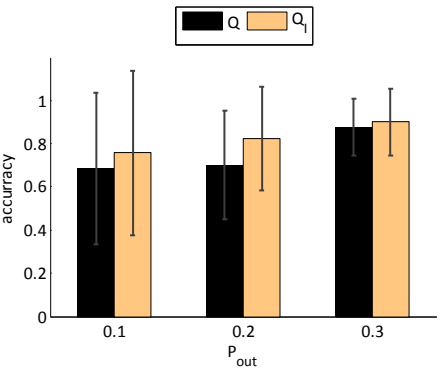

Figure 12: Effectiveness of Q_I: Determining the cluster numbers by the relational modularity Q_I performs better than the modularity function Q, against different degree of community noise (p_{out}).

clustering number C available from the simulation. We quantify the effectiveness by *accuracy* – defined as the portion of instances where the clustering numbers are correctly identified by either Q_I or Q. The experiments are repeated 30 times under each of the different settings and the average performance results are reported.

Figure 12 shows the mean accuracy of Q_I and Q against different degree of clustering noise given by the out-linking probability p_{out}. We observe that both functions tend to underestimate the clustering numbers with small noise. In our experiments, Q_I always outperforms Q in identifying the correct cluster numbers. This indicates the effectiveness of incorporating additional clustering information in the modularity function, including the soft membership and the goodness of a clustering defined by the factorization objective as discussed in section 5.

APPENDIX C. Detailed Prediction Results

The prediction performance averaged over all groups, with 70% and 90% photos for training, are shown in Table 2 and Table 3, respectively.

Table 2: The average tag prediction performance evaluated by four metrics, S@10, P@10, MRR and NDCG, with 70% photos for training and 30% for testing. We compare our prediction results (denoted by "RSC") with three baseline methods: feature-based (denoted by "Features"), tag-based (denoted by "Tags") and feature/tag (denoted by "F/T") based predictions.

	Features	Tags	F/T	**RSC**
S@10	0.320±0.237	0.712±0.183	0.702±0.189	**0.767±0.142**
P@10	0.050±0.039	0.178±0.068	0.174±0.065	**0.251±0.070**
MRR	0.213±0.090	0.566±0.206	0.560±0.205	**0.628±0.170**
NDCG	0.077±0.054	0.300±0.117	0.295±0.114	**0.400±0.111**

Table 3: The average tag prediction performance evaluated by four metrics, S@10, P@10, MRR and NDCG, with 90% photos for training and 10% for testing.

	Features	Tags	F/T	**RSC**
S@10	0.325±0.231	0.700±0.189	0.693±0.192	**0.755±0.142**
P@10	0.050±0.036	0.176±0.072	0.173±0.069	**0.249±0.072**
MRR	0.213±0.090	0.566±0.206	0.560±0.205	**0.628±0.170**
NDCG	0.076±0.052	0.299±0.126	0.293±0.120	**0.397±0.117**

APPENDIX D. User Study

The user study is design to examine whether the recruited participants found the clustering results, extracted by our algorithm, useful, meaningful, and for which purpose. We ask participants to interact with our prototype system called "GAct" and examine how they react to the system. Figure 13 shows a screenshot of the system.

We recruited a total of 12 participants via emails and word-of-mouth. The participants range in age between early 20s and 40s, with diverse backgrounds ranging from non-technical areas to engineering. All the participants have varied experience with online photo sharing websites, and 5 of them have used the Flickr group pool feature.

We asked participants to explore two Flickr groups and then conducted a semi-structured interview session for gathering their feedback on our system. To elicit feedback about group photo exploration, we asked questions such as "Do you find typical patterns or common themes in the group?", "What are the patterns and themes about?", "Could you identify a timeframe for a certain theme?", and "Do you find particular users or tags that associated with the patterns or themes you described?" Finally, we asked the participants to respond to a survey questionnaire. The questions and users' responses are listed in Table 5. In addition, we asked users to select cluster number(s) among 3, 5 and 7

Figure 13: User study. We developed an interactive prototype system "GAct" that generates relational clusters automatically from a given Flickr photo pool and allows users to explore the relationship among photos, users, tags and times. By using the proposed joint factorization method, the system extracts five clusters in this group, retrieves representative users and tags, and renders their membership with each cluster by links, where the link thickness indicates the membership weight. It also retrieves representative photos and generates a timeline for each cluster.

for each group which they thought appropriate for capturing the major themes in the groups. Table 4 shows the user study process.

Table 4: User study process.

Session 1	(a) Explore the "AME" group using the Flickr group Webpages.	(b) Explore the "AME" group using GAct and compare their observation with session 1(a). Compare three versions (3,5, and 7) of relational clustering.
Session 2	(a) Explore the "Antiques" group using GAct and compare the three versions (3,5, and 7) of relational clustering.	(b) Explore the "Antiques" group using Flickr Webpages and compare their observation with session 2(b).
Session 3	Answer a set of survey questions.	

The two groups given to the participants are the "AME"[6] and the "Antiques and their houses"[7] (we use "Antiques" to denote this group in the following description). Note that

[6] http://www.flickr.com/groups/83713445@N00
[7] http://www.flickr.com/groups/20181527@N00

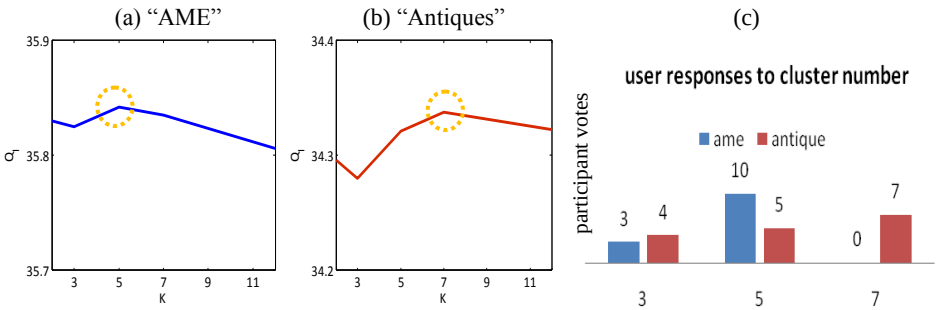

Figure 14: The best clustering numbers determined based on the maximal relational modularity Q_I for the group (a) “AME” and (b) “Antiques” are 5 and 7, which well correspond to participants’ perception shown in (c). The bar chart in (c) shows the number of participants who prefer the corresponding number of themes for the two group content shown in GActs.

8 of the participants are students of the School of Arts, Media and Engineering (AME)[8]. One participant is an active member in the “AME” group. The other 7 participants are familiar with some of the members in the “AME” group, or with events or activities captured by the photos in this group. None of the participants have seen the “Antique” group and none of them has particular interest or expertise in antiques.

Discussion. We report the major positive and negative feedback about the system. Broadly speaking, the participants found the relational clustering results provided by the GAct system clearly represented major themes in a group, the clustering results reflected how participants described the group data, the timeline information of themes were particularly useful for discovering the evolution of the group activity. The major comments are summarized below.

Structure exploration with GAct. Almost all participants felt that GAct provided useful and meaningful structure for the group photos and is particularly helpful for exploring unfamiliar groups such as “Antiques”. With the “AME” group, participants (including those not familiar with the AME School) identified main themes using similar terms “project work,” “party or social events,” and “trips” by simply browsing the Flickr Webpages. When interacting with GAct, they found the 3-cluster results corresponded well to their raw observations. However, participants also agreed that the 5-cluster and 7-cluster results made sense and commented that the versions identified themes in more details (e.g., two themes are similarly understood as “trips” or “landscape” but from different Flickr users, and another two themes are different “project work”). The participants’ exploration of “Antique” group photos was quite different from the “AME” photos. With the “Antiques” group, almost all participants felt overwhelmed when browsing the Flickr Webpages, but found the themes generated by GAct helped guide their exploration.

Structure properties. When interacting with GAct, participants tended to use visual similarity to judge the meaningfulness of a theme, and described the themes based on the visual content depicted in the cluster photos. Participants commented that the timeline information was useful. Participant 3 thought the timeline was important and useful when the theme was about human activity (e.g. projects, parties and trips), but not useful when the theme wais about inanimate objects (e.g. antiques). Participant 6 expressed that the temporal strength of the “party” theme diminished in recent year which was an interesting observation. When exploring the “AME” group, participants who were familiar with the AME School were more interested in playing with the user-theme and tag-theme links. Participant 6 and 9 (from the AME School) noticed the tags “amenities”

[8] www.ame.asu.edu

and "arduino" associated with the theme "project work" made sense. Many participants (including one not from the AME School) noticed tags such as "cake" and "birthday" associated with the theme "party" as meaningful descriptors for the group activity. Participant 3 (from the AME School) identified active contributors in the "trip" theme. She also felt that the links between tags and themes helped her understand "canada" as a high frequency tag in the "trip" theme. With the "Antiques" group, almost all participants did not look at the user-theme links. Some participants were able to make sense of the themes through tag-theme links, e.g. participant 1 found the house photos in one theme were about "abandoned" houses. Participant 9 thought the tag-theme links would be more useful for discovering unfamiliar groups.

Structure complexity. We compare participants' preferred cluster numbers with the optimal cluster numbers determined by the relational modularity function Q_I. As shown in Figure 14, the optimal numbers from Q_I are 5 and 7 for the "AME" and "Antiques" group photos, which correspond well to participants' perception. For the "AME" group, participants said they prefer GAct to merge the 7-cluster results when the themes have similar semantics (e.g. different themes for "project work" can be merged together). Participants 2 and 6 told the interviewer that whether more or fewer cluster numbers would be more useful depended on their purpose and the content of the group – they would prefer 3-cluster if they wanted a quick summary of the group, but 7-cluster if they wanted to find more interesting photos from the group. Participant 2 also noted that, for some groups representing her strongest interests, she thought she would miss many interesting photos in GAct. For the "Antiques" group, many participants felt the 7-cluster results were better than the 5-cluster. Some participants (e.g. 8 and 12) had strong preference for fewer clusters regardless of the group content.

Other consideration. Participants found the interaction in GAct fairly simple. It provided an interesting starting point to explore photos in a group. However, when cluster number increases, the theme panels and links among panels were cluttered and make it difficult to explore more interesting results. Many participants gave valuable suggestions to improve the visual and interaction design of the GAct system, e.g. using icon size instead of line thickness to indicate users' contribution to themes.

Table 5: User survey results from 12 participants, summarized by mean and standard deviation. All ratings are on a 5 point scale (5 is the best and 1 is the worst).

Does GAct help you discover typical patterns or common theme in the shared photos?	4.09±0.54
Does GAct help you discover who contributes most to certain themes?	3.83±1.19
Does GAct help you discover how different or similar tags are associated with certain themes?	3.80±0.79
Does GAct help you discover how visual themes or patterns grow or diminish over time?	3.92±1.24
How much do you prefer GAct to the existing Flickr group navigation?	4.30±0.48
Are photos on a common topic?	4.00±0.60
How much do you think GAct is helpful for understanding group activity?	4.17±0.58
How much do you like GAct?	4.00±0.60

To conclude, the clustering results, including the cluster content and the cluster numbers, of both groups displayed in GAct are deemed meaningful. However, in an exploratory system, the preferred number of clusters may depend on various factors, including the group topics, users' informational need and information processing capacity, the visual design of the system, etc. These factors may not be available from group photo data and need to be further considered in the context of applications. Nevertheless, our framework is able to give reasonable results when such information is not available.