Producing a Unified Graph Representation from Multiple Social Network Views

Derek Greene
School of Computer Science & Informatics
University College Dublin, Ireland
derek.greene@ucd.ie

Pádraig Cunningham
School of Computer Science & Informatics
University College Dublin, Ireland
padraig.cunningham@ucd.ie

ABSTRACT

In many social networks, several different link relations will exist between the same set of users. Additionally, attribute or textual information will be associated with those users, such as demographic details or user-generated content. For many data analysis tasks, such as community finding and data visualisation, the provision of multiple heterogeneous types of user data makes the analysis process more complex. We propose an unsupervised method for integrating multiple data views to produce a single unified graph representation, based on the combination of the k-nearest neighbour sets for users derived from each view. These views can be either relation-based or feature-based. The proposed method is evaluated on a number of annotated multi-view Twitter datasets, where it is shown to support the discovery of the underlying community structure in the data.

ACM Classification Keywords

H.2.8 Database Applications: Data Mining

Author Keywords

Social network analysis, Data integration, Social media

INTRODUCTION

Social networks are often represented using multiple *views* or relations that share all, or part of the same user set. In many cases, these views will consist of graphs with heterogeneous edge types, where each type has different semantics, along with different frequency or weight distributions [2]. For instance, in the case of Twitter, we can characterise users by the accounts whom they follow (or who follow them), the users whom they retweet (or who retweet them), the curated lists to which they have been assigned, and so on. Additionally, users in real-world social networks often have associated attribute information, such as demographic details or user-generated textual content (*e.g.* the content of a user's tweets on Twitter).

For many social network analysis tasks, it will be preferable to work with a unified representation that summarises the information provided by all the data views, rather than working separately on individual views. A variety of community finding algorithms have been proposed in the literature that assume the existence of a single relation between nodes. However, increasingly there is interest in uncovering community structure from richer data sources that provide multiple relations [8]. From a visualisation perspective, it is much easier to interpret a graph with a single aggregated relation (as shown in Fig. 2) than it is to interpret representations that include multiple different types of relations (*e.g.* retweet, follows, mentions). In the task of user curation on social media platforms, it is necessary to combine information from multiple views to produce a definitive set of recommendations [6].

We propose a new method for integrating multiple data views to provide a sparse, unified graph representation, which retains the most informative connections from the original views. The aggregation process is performed at a local level, by combining the ranked neighbour sets for each individual user, and then constructing an overall directed nearest neighbour graph from the local neighbour sets. Unlike many alternative approaches, the views can be either relation-based or feature-based, once a similarity or ranking measure is defined on those views. The views can be incomplete, once there is a partial mapping between the views. Also, there is no requirement to manually select parameters indicating the relative importance of the different views, and no requirement for supervision in the form of labelled training examples. We present evaluations on a set of annotated Twitter datasets, which show that the unified graphs facilitate the identification of meaningful community structure from multi-view data.

RELATED WORK

A range of techniques have been described for clustering across multiple feature-based views. Zhou & Burges [10] proposed a spectral clustering approach for application to multiple graphs sharing the same set of nodes, based on a mixture of Markov chains defined on the different views. The relative importance of each graph is defined by a manually-specified parameter. Greene & Cunningham [5] proposed a "late integration" strategy for clustering heterogeneous data sources, based on the concept of cumulative voting in unsupervised ensembles. The strategy was applied to bibliographic data, consisting of co-citation relations and paper abstracts represented using a bag-of-words model. More recently, Liu *et al.* [7] proposed a joint non-negative matrix factorisation algorithm, which applies an iterative update procedure to find a consensus between the input matrices. The influence of each view on the outcome is determined by a user-specified set of regularisation parameters.

WebSci'13, May 1 – May 5, 2013, Paris, France.
ACM 978-1-4503-1889-1.

In the context of network analysis, the direct integration of multiple relation types can prove difficult, if the relations in the different views are not comparable [8]. Cai *et al.* [2] proposed a regression-based technique to find the optimal linear combination of a number of different weighted relation matrices, relying on a set of input examples that have been assigned community labels. Based on the combined relations, the authors then applied a spectral clustering algorithm to produce disjoint communities.

While most community finding algorithms assume the existence of only one kind of relation, Tang *et al.* [8] focused on the problem of finding groups of related users in "multi-dimensional networks". The authors described a range of alternative strategies, including modularity-based community finding applied to the average interaction network among a group of users, and a "feature integration" strategy where structural features from different views are mapped into the same space.

METHODS

We propose a method to produce a unified network representation from either feature-based or relational views of a set of social network users, based on the application of SVD rank aggregation [9] to a matrix encoding multiple nearest neighbour sets for each user. The per-user rankings are then combined to form a global graph covering all users. This sparse graph represents a unified summarisation of the strongest connections between users across all views.

Neighbour Set Identification

The input to the aggregation process is a dataset of users $\{u_1, \ldots, u_n\}$, along with l different views, each representing some or all of the n users. These views may be relation-based or feature-based. The only requirement is that some measure of similarity is provided for each view – either a metric or non-metric measure can be used. The only parameter required for the aggregation process is a value for the number of nearest neighbours k. This value controls the sparsity of the output graph – a lower value of k will result in a less dense graph. The first phase of the aggregation process is as follows, for each user u_i:

1. For each view $j = 1$ *to* l, compute a similarity vector v_{ij} between u_i and all other users present in that view, using the similarity measure provided for the view.
2. From the values in v_{ij}, produce a rank vector of all other $(n-1)$ users relative to u_i, denoted r_{ij}. In cases where not all users are present in view j, missing users are assigned a rank of $(n'_j + 1)$, where n'_j is the number of users present in the view.
3. Stack all l rank vectors as columns, to form the $(n-1) \times l$ rank matrix $\mathbf{R}_i$, and normalise the columns of this matrix to unit length.
4. Compute the SVD of $\mathbf{R}_i^{\mathsf{T}}$, and extract the first left singular vector. Arrange the entries in this vector in descending order, to produce a ranking of all other $(n-1)$ users. Select the k highest ranked users as the *neighbour set* of u_i.

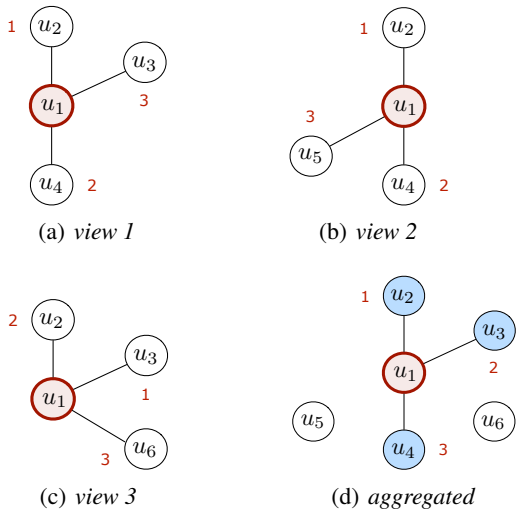


Figure 1. Example of the proposed aggregation method, involving six users and three views. Graphs (a)-(c) show the ranked neighbour sets for the user u_1 for $k = 3$. By combining the ranks from these neighbourhoods, we produce the aggregated neighbour set $\{u_2, u_3, u_4\}$ for u_1, as shown in (d).

A simple example illustrating the method is shown in Fig. 1. The procedure can be readily parallelised by processing multiple users simultaneously. In addition, the time required for the aggregation process can be reduced considerably by computing the truncated SVD of the rank matrices.

Unified Graph Construction

Once the k-nearest neighbour sets have been identified for all n users, we use this information to build a global graph representation of the dataset. A natural approach to combine the sets is to construct the corresponding asymmetric *k-nearest-neighbour graph*. Specifically, we construct a directed unweighted graph, where each node is a user and an edge exists from node i to j, if u_j is contained in the neighbour set of u_i. This process yields a sparse, unified graph encoding the connectivity information derived from all original views in the dataset, representing all users that were present in one or more of those views.

EVALUATION

We empirically examine the degree to which the proposed aggregation method preserves the most informative underlying associations between users in the original views. For our evaluation, we use four Twitter datasets [4] for which sets of manually-curated ground truth communities are available.

- *football*: A collection of 248 English Premier League football players and clubs active on Twitter. The disjoint ground truth communities correspond to the 20 individual clubs in the league.
- *olympics*: A dataset of 464 users, covering athletes and organisations that were involved in the London 2012 Summer Olympics. The disjoint ground truth communities correspond to 28 different sports.

- *politics-uk*: 419 Members of Parliament (MPs) from the United Kingdom. The ground truth consists of five groups, corresponding to political parties.
- *rugby*: A collection of 854 international Rugby Union players, clubs, and organisations active on Twitter. The ground truth consists of communities corresponding to 15 countries. The communities are overlapping, as players can be assigned to both their home nation and the nation in which they play club rugby.

For each dataset, we constructed a heterogeneous collection of views, containing some or all of the complete set of Twitter users for that dataset.

- **Content views:** tweet content profiles (500 most recent tweets), user list text (merged names and descriptions for 500 most recent lists).
- **Network views:** follows, followed-by, mentions, mentioned-by, retweets, retweeted-by, co-listed.

In all cases, cosine similarity is applied to compute pairwise similarities. For a more detailed explanation of the construction of the views listed above, consult [6].

Evaluation Measure

The concept of *k-nearest-neighbour consistency* for clustering was formalised by Ding & He [3]: for any item in a cluster, its k nearest neighbours should also be assigned to that cluster. Motivated by this work, we evaluate the degree to which alternative representations of a dataset preserve the k-nearest-neighbour consistency of a set of ground truth communities. A representation that reflects the ground truth will have a high level of consistency, while a representation that does not preserve the structure of the ground truth will yield a low level of consistency.

For a single user u_i and view, we can compute the *user consistency* as the fraction of that user's k nearest neighbours in that view that are assigned to the same ground truth community. In the case of overlapping ground truth communities, we generalise by counting the fraction of neighbours that are assigned to at least one community also containing u_i. We then compute the overall *average consistency* as the simple average of all n user consistency scores.

Discussion

When we apply our proposed approach to the four Twitter datasets, a visual inspection of the output (based on force-directed layouts produced using Gephi [1]) highlights the sparsity of the unified graphs. This often results in almost entirely disconnected components, where users assigned to the same ground truth communities are densely-connected, while there is little connectivity between those communities. As an example, Fig. 2 shows the layout for the unified *politics-uk* graph ($k = 5$). We see that there is a clear separation between the various political groupings, with only a handful of long-range inter-community links. Given that we can see the separation clearly by visual inspection alone, any reasonable single-mode community finding algorithm should be able to identify this grouping.

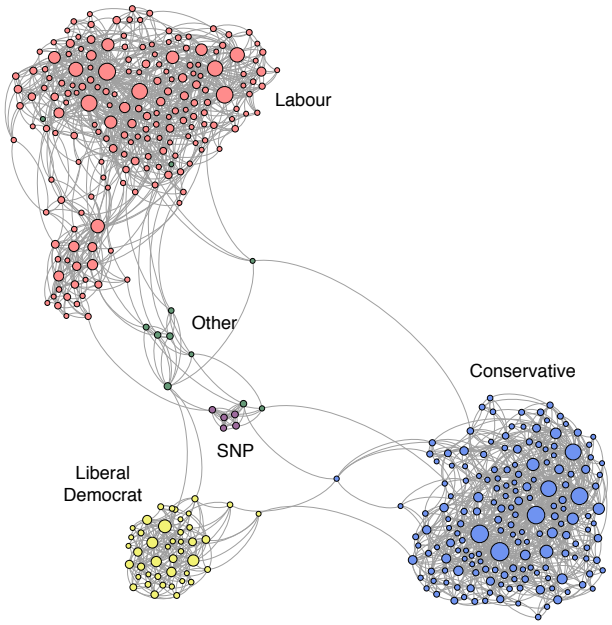
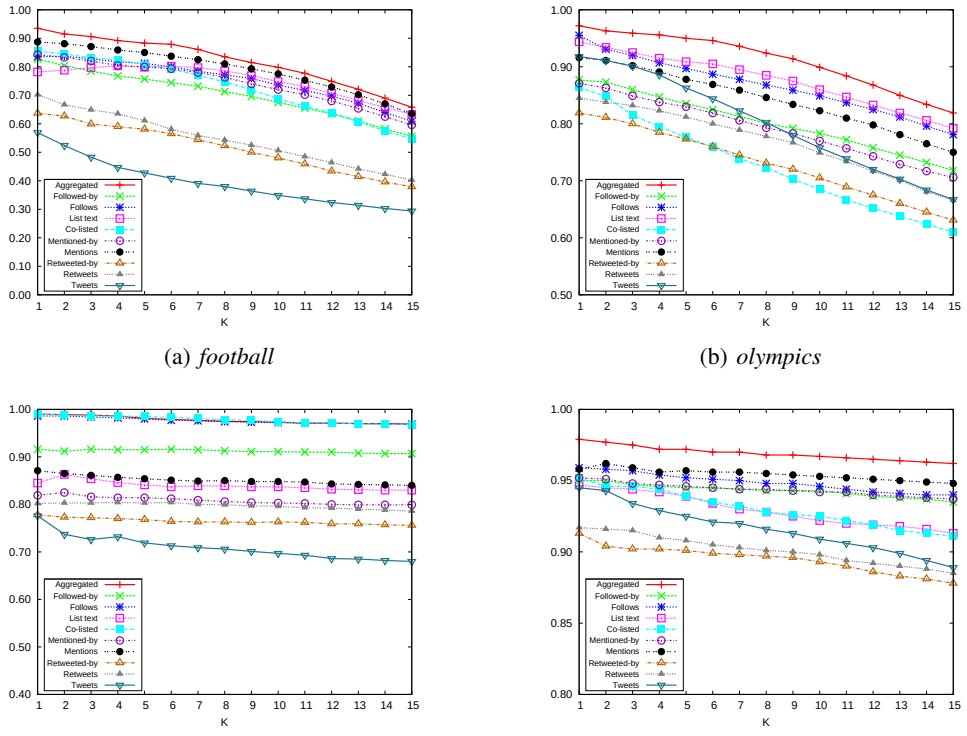


Figure 2. Unified graph generated on the *politics-uk* dataset ($k = 5$). Users are coloured and labelled based on a ground truth, corresponding to five different political groupings.

It is interesting to note that our method also supports the discovery of sub-communities relative to the ground truth, which had not been identified manually. In Fig. 2, we observe that the community for the Labour Party contains a smaller sub-community of users. On inspecting these accounts, it is apparent that they correspond to Labour Party MPs based in Scotland. Similarly, for the *rugby* dataset, we see sub-groups corresponding to individual clubs based in England, Ireland, Scotland, and Australia.

To quantitatively analyse the effectiveness of our method, we compared the average consistency scores for the k-nearest neighbours in the individual views with those achieved by the unified graph, for neighbourhood sizes $k \in [2, 15]$. Fig. 3 shows that, in three of the four datasets, the unified graph provides a higher level of consistency among neighbours than any of the individual views. It is only in the case of the *politics-uk* dataset that the co-listed view out-performs the aggregated approach. Here it appears that there is a high number of carefully-curated user lists on Twitter corresponding to UK political party memberships, while the aggregated approach is somewhat affected by noisy tweet content.

One key observation that can be made is that no single individual view out-performs all others on every dataset. So while, for example, user lists are highly-informative in the case of the *politics-uk* dataset, they prove less useful for identifying distinct groups in the case of the *olympics* dataset. In general, we will typically not know *a priori* which view is most informative. This will be problematic for methods that require the relative importance of each view to be specified as an input parameter. In contrast, our proposed method does not require manual weighting of the views, and performed robustly across all datasets in our experiments.



(a) *football* (b) *olympics*

(c) *politics-uk* (d) *rugby*

Figure 3. Comparison of *average consistency* scores for $k \in [2, 15]$, calculated on nine individual views and the resulting unified graph, across four Twitter datasets.

CONCLUSIONS

We have demonstrated that we can use a form of rank aggregation applied to nearest neighbour sets to construct a single unified graph from multiple heterogeneous data views. Evaluations on annotated Twitter datasets have shown that the unified graphs highlight the underlying community structure. We suggest that this procedure will prove useful as a step prior to other network analysis tasks, such as community finding, visualisation, and user recommendation. Currently, each user in the unified graph has at most k outgoing edges. We plan to examine the adaptive selection of k on a per-user basis, to allow for hubs with many connections, or outliers with few connections.

Acknowledgments. This research was supported by Science Foundation Ireland Grant 08/SRC/I1407 (Clique SRC).

REFERENCES

1. M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proc. 3rd Int. Conf. on Weblogs and Social Media (ICWSM'09)*, pages 361–362, 2009.
2. D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining hidden community in heterogeneous social networks. In *Proc. 3rd International Workshop on Link Discovery*, pages 58–65, 2005.
3. C. Ding and X. He. K-nearest-neighbor consistency in data clustering: Incorporating local information into global optimization. In *Proc. ACM Symposium on Applied Computing (SAC'04)*, pages 584–589, 2004.
4. D. Greene. Twitter Network Datasets, 2013. `http://mlg.ucd.ie/networks`.
5. D. Greene and P. Cunningham. Multi-view clustering for mining heterogeneous social network data. In *Proc. Workshop on Information Retrieval over Social Networks, ECIR'09*, 2009.
6. D. Greene, G. Sheridan, B. Smyth, and P. Cunningham. Aggregating Content and Network Information to Curate Twitter User Lists. In *Proc. 4th ACM RecSys Workshop on Recommender Systems & The Social Web*, 2012.
7. J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. SIAM Data Mining Conf. (SDM'13)*, 2013.
8. L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33, 2012.
9. G. Wu, D. Greene, and P. Cunningham. Merging multiple criteria to identify suspicious reviews. In *Proc. 4th ACM Conference on Recommender Systems (RecSys'10)*, pages 241–244. ACM, 2010.
10. D. Zhou and C. Burges. Spectral clustering and transductive learning with multiple views. In *Proc. 24th Int. Conf. Machine Learning*, 2007.