Covering the Egonet: A Crowdsourcing Approach to Social Circle Discovery on Twitter

Karmen Dykstra[⋆] and **Jefrey Lijffijt**[⋆†] and **Aristides Gionis**[⋆]
[⋆] Helsinki Institute for Information Technology (HIIT)
Department of Information and Computer Science
Aalto University, Finland
[†] Intelligent Systems Laboratory
Department of Engineering Mathematics
University of Bristol, United Kingdom

Abstract

Twitter and other social media provide the functionality of manually grouping users into *lists*. The goal is to enable selective viewing of content and easier information acquisition. However, creating lists manually requires significant time and effort. To mitigate this effort, a number of recent methods attempt to create lists automatically using content and/or network structure, but results are far from perfect.

In this work, we study the power of the millions of lists that are already created by other twitter users in order to "crowdsource" the task of list creation. We find that in a large dataset, collected specifically for this study, an optimal matching of existing lists from other twitter users to the ground-truth lists in egonets gives an F_1 score of 0.43, while the best existing method achieves only 0.21.

We explore the informativeness of features derived from network structure, existing lists, and posted content. We observe that different types of features are informative for different lists, and introduce a simple algorithm for ranking candidate lists. The proposed algorithm outperforms existing methods, but still falls short of the optimal selection of existing lists.

Introduction

Modern social networks allow users to organize their friends into groups — social circles on Google+, community pages on Facebook, and lists on Twitter. This functionality allows users to selectively interact with posted content, and more easily browse relevant information. However, grouping users manually is a time-consuming task, and it is thus desirable to automate the group-creation task.

The task is formally known as the *social-circle discovery* problem (McAuley and Leskovec 2012). In this problem, the unit of analysis is the *egonet*, which is the local network formed by a user's friends and the friendship links between them. The goal is to discover the most meaningful clusters of users in the egonet. Performance is typically measured by comparing discovered clusters with ground-truth groupings that users have already created in their egonets.

In this work, we study the social circle discovery problem in twitter, in which groups are known as *lists*. Any twitter user can create a list by manually selecting a subset of twitter users and designating a list title. If a user makes a list *public*, then other users can *follow* that list, as well. It has been shown that lists are used to group users who share common expertise, have common interests, belong to the same group of friends, or are located in the same geographical region. Lists are also used to aggregate updates on breaking news stories (Morrison and Hayes 2012; Greene et al. 2012).

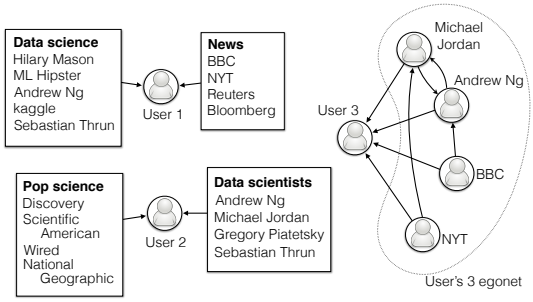

Figure 1: Users 1 and 2 have created some twitter lists, while user 3 has not. We want to utilize the additional information provided by users 1 and 2, to help user 3 organize her egonet into lists. In this example, boldface text indicates list titles.

Previous work on creating twitter lists automatically has achieved moderate success, with one method achieving F_1 score of 0.33, on a dataset collected for their study (Yang, McAuley, and Leskovec 2013), and another method achieving normalized mutual information of 0.27, on a different dataset (Morrison and Hayes 2012). Both of these methods rely on features derived from *tweet content* and from the *structure* of the follower graph.

Our premise is that these methods ignore an important source of information: the millions of lists *already created* by twitter users. The information value of these lists is multifaceted; co-listing of two users suggests similarity between them, while list titles have been shown to indicate user interests (Kim et al. 2010). At the same time, existing lists cover a large portion of twitter users; in our dataset 87% of the users belongs to at least one list. On the other hand,

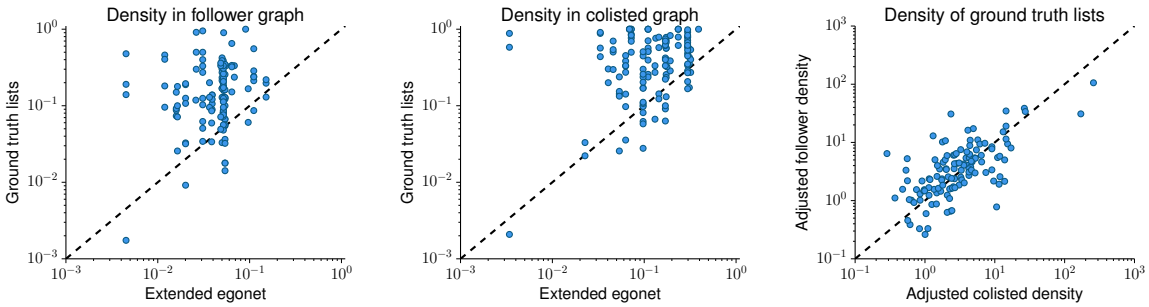

Figure 2: Connection density in lists vs. full egonets for the follower graph and the co-listed graph, and a comparison that shows the correlation between the two measures (note log-scale axes).

note that the problem we consider is different than the list-recommendation problem (Rakesh et al. 2014), since our goal is to group the friends of the egonet of a user into meaningful lists, and not to recommend any relevant list to a user.

Contributions. We introduce the use of *existing* twitter lists for "crowdsourcing" the task of automatically organizing the friends of a user into lists. We use the existing lists that overlap with each egonet as candidate lists. Since the number of such lists can exceed 1 million, we define an optimization problem for selecting k high-quality candidate lists based on an empirical analysis of list and egonet properties.

We prove that the select-k problem is **NP**-hard, but the objective function is submodular, enabling a greedy algorithm that provides a good approximation of the optimal solution. To evaluate the performance of our method, we collect a dataset of 24 egonets with all necessary information. We compare our method with CESNA, a state-of-the-art overlapping-clustering method (Yang, McAuley, and Leskovec 2013). We find that our method selects candidates with titles similar to the ground-truth and that results are on-par with the comparison method, while there is also room for improvement.

Background

Terminology: In Twitter, if user a follows user b, then a is a *follower* of b, and b is a *friend* of a. We work with the *extended egonet*, defined as the friends of the ego user, the members of the lists that the ego subscribes to or owns, and the follower relationships between them. We include such list members as the ego user indirectly follows those users via interaction with the lists.

Data collection: We collect 24 extended egonets using the public twitter API. Ego users are selected using a random walk starting from Stephen Colbert, a prominent comedian with almost 7 million followers. A visited user is selected if (i) they subscribe to or own at least 3 lists and (ii) the size of the union of their friends and followers is less than 3 500 users. To help avoid collection of spam and business accounts we also ensure that the ratio of friends to followers, and vice versa, does not exceed 3:1. We manually check egonet pages in twitter to ensure that they are personal accounts.

For every user in the egonet, we collect (i) list memberships and titles; (ii) hashtags and mentions from the 200 most recent tweets; and (iii) follower relationships between users in the egonet.

In total the dataset includes 42 946 users, 4 181 698 follower relationships, and 25 699 928 list memberships. To evaluate the quality of our methods, we use all 135 lists that our ego users subscribe to or own that contain at least three members, henceforward referred to as *ground-truth lists*.

Distinguishing features

As we will see below, current methods based on tweet content and network structure do not achieve high scores on retrieving the ground-truth lists. Before introducing another algorithm to solve the task, we first study in detail the features that intuitively provide useful information.

Density of lists: Previous work has found that users within lists are densely connected by follower relationships (Morrison and Hayes 2012). Besides density in the follower graph, we expect that being co-listed is a strong indicator of user relatedness. To test this hypothesis, we form the co-listed graph, where there is an edge between users if they already appear together in a list (excluding the considered list). We then examine the density of lists in the follower graph and the co-listed graph, relative to the density of the entire extended egonet. We measure density by $\frac{|E|}{|V||V-1|}$.

The results are showed in Figure 2 (left and middle). We observe that most lists have high density in the follower graph, compared to the extended egonet. The same is true for the co-listed graph. We also observe that there are many ground-truth lists whose density is substantially lower than the density of the egonet. Obviously, an algorithm based only on density cannot identify such lists.

The rightmost plot of Figure 2 shows follower and co-listed density divided by the density of the egonet. We observe that they are correlated, but also that for some lists both ratios are below 1, indicating that neither is informative, as well as that there exist lists for which follower density is informative but not co-listed density and vice versa. Thus, existing lists made by other users provide information that cannot be found by looking only at the network structure.

Topical vs. social lists: It has been argued that communities

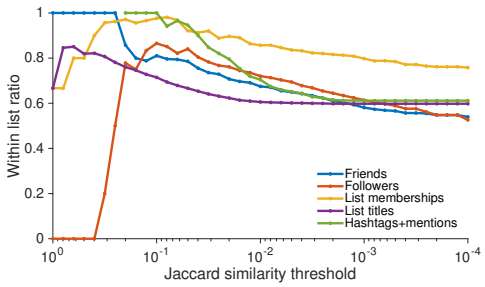

Figure 3: Ratio of number of pairs within lists vs. an equal-sized random sample of pairs within an egonet.

in social media platforms primarily form either around topical or around social relationships (Grabowicz et al. 2013). To gain further insight in the informativeness of network structure, tweet content, and existing lists, we compute the following features for pairs of users, both in lists and in egonets: Jaccard similarity of: (1) friends (2) followers (3) list memberships (4) list titles, and (5) hashtags and mentions in their tweets.

Figure 3 presents ratios of pairs within lists vs. within egonets for each of these features. Each is informative, while we find that having many of the same friends indicates high relatedness, and being co-listed is informative even if both are part of many lists (Jaccard small but > 0). Finally, we analyze these features also for all ground-truth lists individually, some examples of which are given in Figure 4.

Method

Our method is a greedy algorithm that, given an egonet and a set of available lists, selects k of those lists that provide a good coverage of the egonet; "good" will be defined shortly. Motivated by the high density of the ground-truth lists, we apply a density-based quality metric. The method allows for list overlap, as it happens with ground-truth lists.

Problem formulation: We are given a set of users $V = \{v_1, \ldots, v_n\}$ in a target egonet, and a collection of subsets of users, or candidate lists, $\mathcal{L} = \{L_1, \ldots, L_m\}$, $L_j \subseteq V$. These candidate lists are obtained by all existing lists that include at least three users in the egonet. Given a function $s(\cdot)$ that assigns a score to each v_i, our goal is to select k lists that maximize the summed score over all users. The objective can be formally defined as follows:

$$\text{maximize} \quad f(\mathcal{L}) = \sum_{v \in V}^{n} s(v) \quad \text{such that} \quad |\mathcal{L}| = k.$$

We also define a quality function over the lists $q : \mathcal{L} \to [0, 1]$, such that $q(L)$ measures to what degree L exhibits desirable properties. The score of an item $s(v)$ is then defined as the quality of the best list in which v is a member, i.e.,

$$s(v) = \max_{L \ni v | L \in \mathcal{L}} q(L). \tag{1}$$

We experiment with four different q functions. Motivated by the analysis presented in the previous section, we define the quality of the list as the density of its members in the friendship or co-listed graph. For density we use both average degree $q(L) = |E|/|V|$, and normalized density $q(L) = |E|/(|V||V-1|)$.

Table 1: F_1 score averaged across ground truth lists. G/X/Y denotes different variants of the greedy method. X=F: friendship graph; X=C: co-listed graph; Y=A: average degree density; Y=N: normalized density.

	G/F/N	G/F/A	G/C/N	G/C/A	CESNA	BEST
$k = 3$	.142	.168	.156	.153	.187	.539
$k = 20$	.200	.186	.221	.196	.213	.426

Table 2: Titles and sizes of ground truth lists and existing lists selected by our objective function for the top 10 matches. Asian character titles (**) are not displayed.

Ground truth title	size	Greedy title	size	F_1
streetmeets	3	food-trucks-nyc	3	1.00
best-book-review-feeds	82	book-reviews	82	1.00
rebeldes	39	rebelde	33	0.92
entertainment	34	entertainer	27	0.85
aerco	46	grupos-aerco	43	0.74
** [Christianity]	27	truth-chaser	27	0.63
marintweetup	58	bay-area	39	0.62
**	44	**	21	0.62
alimales	7	alimales	13	0.60
**	12	**	5	0.59

Algorithm: We can show that the problem defined above is **NP**-hard, while the objective function f is nondecreasing and submodular, as long as the function s is also nondecreasing and submodular. The latter fact is a consequence of the non-negativity of the density functions we consider and the max function in Equation (1). The proof follows the lines of the proof of submodularity for the uncapacitated facility location problem (Cornuejols, Fisher, and Nemhauser 1977). We omit the details for brevity. As a result of the monotonicity and the submodularity of f, the greedy algorithm can be shown to achieve a $\frac{e-1}{e}$ approximation guarantee.

Experiments

Our problem definition takes as input an egonet, a set of candidate lists, and a number k, defined as the number of lists to be selected. We experiment with $k \in \{3, 20\}$, where 3 reflects a use-case scenario in which we want to select the three best lists for a user, and 20 is the maximum number of lists in any egonet in our dataset. As list candidates we input the existing lists to which at least 3 egonet users belong.

Our method also takes as input the function q that specifies the quality of each candidate list. We experiment with four different settings, which we compactly denote by G/X/Y. G stands for Greedy. X $\in \{$F, C$\}$ indicates whether we are using the friendship graph or the co-listed graph. Y $\in \{$N, A$\}$ indicates whether we are using the normalised density or the average-degree density.

To evaluate the predicted lists, we first find the linear assignment of ground truth lists to predicted lists that maximizes the summed F_1 score in each egonet. That is, given ground truth lists $\mathcal{L}_{\text{GT}}$ and predicted lists $\mathcal{L}_{\text{P}}$, we find a map-

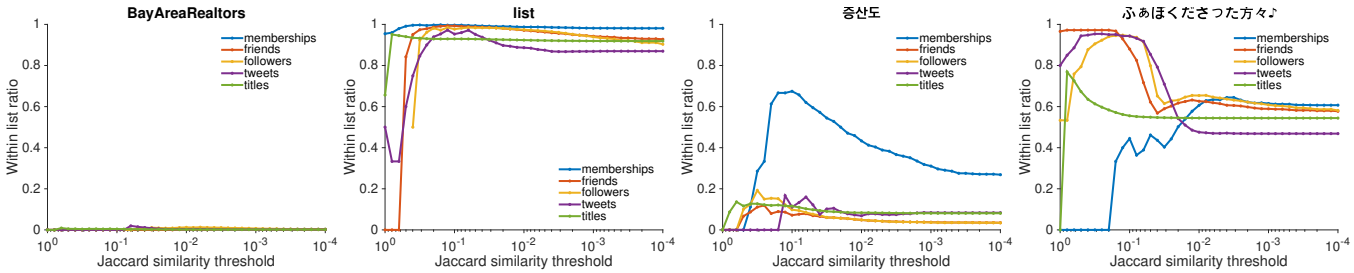

Figure 4: Canonical examples of feature informativeness per list; sometimes no feature puts pairs from the list ahead of the rest of the egonet, while in other instances all features work. However, in many cases, either one or a few features are somewhat informative, there is no consistent pattern that would strictly favor one feature over another.

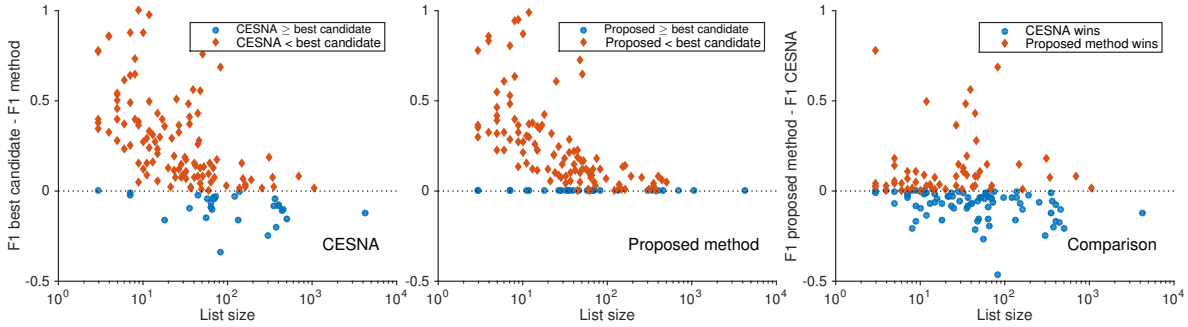

Figure 5: Ground truth list size versus difference in F_1 score between **left**: BEST and CESNA **middle**: BEST and G/C/N and **right**: CESNA and G/C/N.

ping $f : \mathcal{L}_{\text{GT}} \rightarrow \mathcal{L}_{\text{P}}$ that maximizes $\sum_{L \in \mathcal{L}_{\text{GT}}} F_1(L, f(L))$.

We also calculate an upper bound on the F_1 score that can be achieved by using only existing lists as selection candidates. When $k = 20$ we select the existing list that maximizes the F_1 score averaged over ground truth lists. When $k = 3$ we select the 3 pairs of existing lists and ground truth lists that yield the highest F_1 score. This method is denoted as BEST in Table 1. For comparing to CESNA, we use the default parameters provided in the SNAP package.[1]

Results: Our results are shown in Table 1 and Figure 5. The proposed method achieves a highest F_1 score of .22 when the greedy method is applied in conjunction with the normalized density of lists in the colisted graph. Normalized density achieves higher results than average degree in the co-listed graph, and vice versa in the follower graph. Results are on-par with the .21 achieved by CESNA. As seen in the far-right plot in Figure 5, CESNA dominates G/C/N on large lists. A qualitative comparison of the lists found with our method, against ground truth lists, is shown in Table 2, in terms of titles, sizes, and F_1 scores.

Discussion

We introduce a method for using existing lists on twitter for recommending informative groupings in a user's local network. The method is shown to be on-par with the state-of-the-art overlapping clustering algorithm, while our analysis demonstrates that there is significant room for improvement on the task. One shortcoming of the approach is its inability to create lists for users who do not already belong to a list outside of the ground-truth. A possible solution to this would be the application of a seed set expansion method.

[1] http://snap.stanford.edu/snap/

References

Cornuejols, G.; Fisher, M.; and Nemhauser, G. L. 1977. On the uncapacitated location problem. *Annals of Discrete Mathematics*.

Grabowicz, P. A.; Aiello, L. M.; Eguíluz, V. M.; and Jaimes, A. 2013. Distinguishing topical and social groups based on common identity and bond theory. In *ICWSM*.

Greene, D.; Sheridan, G.; Smyth, B.; and Cunningham, P. 2012. Aggregating Content and Network Information to Curate Twitter User Lists. In *RecSys*.

Kim, D.; Jo, Y.; Moon, I.-C.; and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*.

McAuley, J. J., and Leskovec, J. 2012. Learning to discover social circles in ego networks. In *NIPS*.

Morrison, D., and Hayes, C. 2012. Early and late fusion methods for the automatic creation of twitter lists. In *ASONAM*.

Rakesh, V.; Singh, D.; Vinzamuri, B.; and Reddy, C. K. 2014. Personalized recommendation of twitter lists using content and network information. In *ICWSM*.

Yang, J.; McAuley, J.; and Leskovec, J. 2013. Community detection in networks with node attributes. In *ICDM*.