Analysis of Cluster Structure in Large-scale English Wikipedia Category Networks

Thidawan Klaysri, Trevor Fenner, Oded Lachish,
Mark Levene, and Panagiotis Papapetrou

Department of Computer Science and Information Systems,
Birkbeck, University of London, UK

Abstract. In this paper we propose a framework for analysing the structure of a large-scale social media network, a topic of significant recent interest. Our study is focused on the Wikipedia category network, where nodes correspond to Wikipedia categories and edges connect two nodes if the nodes share at least one common page within the Wikipedia network. Moreover, each edge is given a weight that corresponds to the number of pages shared between the two categories that it connects. We study the structure of category clusters within the three complete English Wikipedia category networks from 2010 to 2012. We observe that category clusters appear in the form of well-connected components that are naturally clustered together. For each dataset we obtain a graph, which we call the *t-filtered* category graph, by retaining just a single edge linking each pair of categories for which the weight of the edge exceeds some specified threshold t. Our framework exploits this graph structure and identifies connected components within the *t-filtered* category graph. We studied the large-scale structural properties of the three Wikipedia category networks using the proposed approach. We found that the number of categories, the number of clusters of size two, and the size of the largest cluster within the graph all appear to follow power laws in the threshold t. Furthermore, for each network we found the value of the threshold t for which increasing the threshold to $t+1$ caused the "giant" largest cluster to diffuse into two or more smaller clusters of significant size and studied the semantics behind this diffusion.

Keywords: graph structure analysis; large-scale social network analysis; Wikipedia category network; connected component

1 Introduction

Wikipedia is one of the most popular large social media networks and has experienced exponential growth in its first few years of existence in terms of articles, page edits, and users [4]. Moreover, this large network has been studied extensively; for example, analysis of the social networks emanating from the Wiki-talk page or discussions page reveals rich social interactions between editors [1, 9, 12]. On the other hand, the Wikipedia network of category links, which indirectly implies social relations when authors assign their articles into specific categories,

has received much less attention from the research community, in particular in terms of large-scale structural social network analysis. Current research on the Wikipedia category network has mainly concentrated on content-based analysis.

The Wikipedia category network mainly consists of categories, where two categories are connected by an edge if they have some "similiarity". In our setting, similarity is expressed by the number of pages shared between two categories. In other words, the weight of an edge is equal to the number of common pages between the categories, and hence expresses the similarity between them: the higher the weight, the higher the similarity.

Wikipedia categorisation refers to assigning an article to at least one category to which it logically belongs. The Wikipedia categorisation system is likely to be improved in the long run, as category policies are still being refined[1]. There is no limit to the size of the categories, but when a category becomes very large, it may be diffused (or broken down) into smaller categories or subcategories. This phenomenon is called *large category diffusion*.

Our objective in this paper is to examine the structural properties of the category clusters within the Wikipedia category network by identifying well-connected components in the graph. These components can be used for comparison with the Wikipedia category tree, based on the expectation that categories falling into same cluster should have a high degree of proximity within the Wikipedia tree.

Our key contributions are summarized as follows:

- We present *t-component*, a framework for identifying natural category clusters in the form of well-connected components in a category-links network that employs an edge-weight threshold t regulating the "strength" of the components.
- Using the proposed framework, we study several structural properties of the Wikipedia network, such as the number of non-trivial category clusters, the size of the largest category cluster, and the number of the smallest category clusters, and how they evolve as the edge-weight threshold increases.
- We observe the diffusion of the largest category cluster as a "giant"-cluster splitting into smaller sub-clusters and examine their contents.
- We find that the largest connected component shrinks at a power-law rate as the edge-weight threshold t increases. This is consistent with similar observations for various properties of social networks, such as the Barabasi-Albert model, which is considered a reasonable generative model of the Web.

2 Related Work

Analysis of web social networks has become a popular research area, especially in the context of online social networking applications. Some large-scale networks have been analysed recently. For example, social interactions have been analysed in Twitter [11], Wattenhofer et al. [19] analysed the nature of the YouTube

[1] http://en.wikipedia.org/wiki/ Wikipedia:FAQ/Categorization

network, Sadilek et al. [13] modelled the spread of diseases by analysing health messages from Twitter, Volkovich et al. [18] analysed structural properties and spatial distances of the Spanish social network Tuenti, and finally Goel et al. [3] studied user browsing behaviour changes.

Wikipedia, which is one of the most popular social media networks has been studied extensively. For example, Hu et al. [4] analysed and predicted user collaborations, Leskovec et al. [10] investigated the promotion process from the point of view of the voters engaged in group decision-making, and Jurgens et al. [6] investigated trends of editor behaviour. The page links structure has also been studied. For instance, Buriol et al. [2] examined the page links structure and its evolution over time and Kamps et al. [7] compared the Wikipedia link structure to other similar web sites. Also, a survey on graph clustering methods by Schaeffer [14] provides a thorough review of different graph cluster definitions and measures for evaluating the quality of clusters.

There were a lot of studies of the Wikipedia user talk pages in the context of its induced social network, which contains rich social interactions in the "talk" domain. Examples include analysing the policy governance discussed on user talk pages [12] and detecting structural patterns forming a tree structure [9].

In general, category links in the Wikipedia category network have been studied using text analysis. For example, Schonhofen [15] attempted to identify document topics, while Kittur et al. [8] represented topic distribution mapping with category structure, and Jiali et al. [5] studied document topic extraction. While Zesch and Gurevych [20] analysed the Wikipedia category graph from a natural language processing perspective, the large-scale Wikipedia network structure has been studied much less. For example, Suchecki et al. [17] investigated the evolution of the English Wikipedia category structure from 2004 to 2008; but, focusing merely on the structure of the documentation of knowledge.

3 Preliminaries

In this paper, we focus on the Wikipedia category network, which we describe next. Then we provide the necessary background definitions to be used in the remainder of this paper.

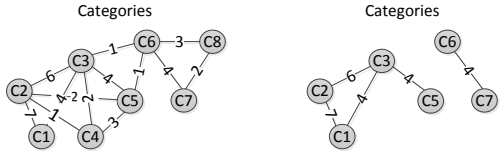
3.1 The Wikipedia Category Network

Wikipedia contains knowledge in the form of Wiki pages and is edited collaboratively by millions of volunteer authors in 285 different languages, among which the English Wikipedia contains the largest number of articles. In Wikipedia, each article is assigned to at least one category, while a categorised article should be assigned to all of the most specific subcategories to which it logically belongs.

Assigning pages to the categories induces a social network of pages and categories established by the editors. This network can be considered as a graph, representing a set of relationships between pages and categories, or only between categories.

3.2 Problem Setting

Let $P = \{p_1, \ldots, p_n\}$ be the set of n Wikipedia pages and C be the set of m Wikipedia categories. Each page $p_i \in P$ belongs to at least one category $c_j \in C$. A graph $G = (V, E)$ is defined as a set of vertices V and a set of edges E, such that each edge $e_k \in E$ connects two vertices $v_i, v_j \in V$, which is denoted as $v_i \rightarrow^{e_k} v_j$.



(a) a category graph (b) the 4-filtered category graph of (a)

Fig. 1: Examples of two category graphs.

Definition 1 `(page-category graph)`
A page-category graph is a bipartite graph G^{PC} that represents the network of connectivity between Wikipedia pages and Wikipedia categories. The set of vertices is $P \cup C$ and there is an edge $p \rightarrow c$ whenever page $p \in P$ belongs to category $c \in C$.

Each page in P belongs to at least one category in C. A page that belongs to a single category is called an *isolated page*.

Definition 2 `(edge-weighted category graph)`
An edge-weighted category graph G^{EW} is a graph where the set of vertices corresponds to the Wikipedia categories C. Each edge e_k between two vertices v_i and v_j (corresponding to categories c_i and c_j, respectively) is assigned with a weight $w_k \in \mathbb{N}$ equal to the number of common pages in both c_i and c_j, or equivalently

$$w_k = |\{p \in P \mid p \in c_i \text{ and } p \in c_j\}|$$

Note that an edge e_k with weight $w_k = 1$ is called *a feeble edge* and a category that is not sharing any page with any other category is called an *isolated* or *trivial category*. It follows that pages connected to an isolated category are necessarily isolated.

Definition 3 `(t-filtered category graph)`
A t-filtered category graph G_t^{EW} is obtained from an edge-weighted category graph G^{EW} by the removal of every edge e_k with weight less than $t \in \mathbb{N}$, i.e., e_k is in G_t^{EW} if and only if

$$w_k \geq t \ , \forall e_k \in E.$$

In Figure 1 we see two examples of category graphs. The first one (on the left) is a category graph where no filtering has been applied, while the second one (on the right) is the corresponding 4-filtered category graph of the one on the left. We note that the t-filtered graph is closely related to the m-core of a graph [16].

Definition 4 `(category cluster)`
A category cluster $\mathcal{C}(G^{EW}, t)$ *is a well-connected component of an edge-weighted category graph* G^{EW}*, and is obtained as a connected component of the corresponding t-filtered category graph* G_t^{EW} *for a specified threshold* $t \geq 2$*.*

Using the above definitions we can now formulate the problem studied in this paper as follows.

Problem 1 *Given a Wikipedia page-category graph* G^{PC} *and a theshold* $t \in \mathbb{N}$*, identify the largest category cluster in the corresponding t-filtered category graph* G^{EW}*.*

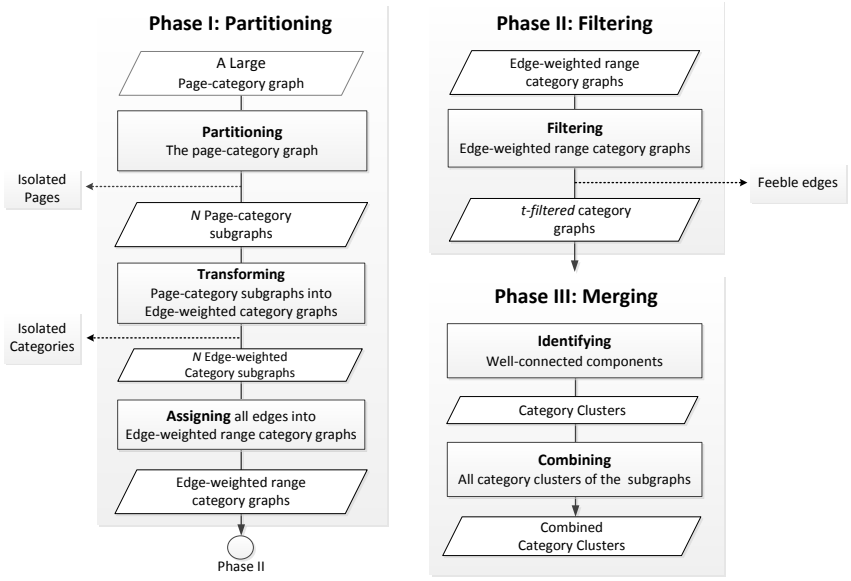


Fig. 2: An overview of the six steps of the t-component framework.

4 The t-component Framework

The *t-component* framework consists of the three main phases as shown in Figure 2: (I) partitioning the edge-weight category graph, (II) filtering the subgraphs, and (III) merging the subgraphs. Next we descibe each phase in more detail.

4.1 Phase I: Partitioning

Due to the difficulty of manipulating the entire page-category graph, which is very large, in this phase, the input bipartite page-category graph G^{PC} is partitioned into a set of N page-category subgraphs $\{G_1^{PC}, \ldots, G_N^{PC}\}$. The split is performed ensuring that: (1) each page $p \in P$ appears in only one subgraph and (2) all subgraphs have approximately the same number of pages. In addition, all isolated pages are eliminated.

Next, each page-category subgraph G_i^{PC} is transformed into its corresponding edge-weighted category subgraph G_i^{EW}, following Definition 2. Note that all isolated categories are eliminated. An important observation here is that, during this process, the same edges may appear in more than one edge-weighted category subgraph. In other words, we could have different sets of pages shared between the same two categories c_i, c_j within different subgraphs. We note that G^{EW} is the union of all edge-weighted category subgraphs G_i^{EW}, for $i = 1, 2, \ldots, N$. Due to its size we cannot explicitly construct G^{EW}.

Therefore, we construct a collection of edge-disjoint category graphs whose union is G^{EW}. Towards this end, we split the initial set of categories C into a set of R category ranges $\mathcal{R} = \{r_1, \ldots, r_R\}$ of approximately equal length, where each range $r_t = [r_t^l, r_t^u)$ defines a lower ($r_t^l \in [|C|]$) and upper ($r_t^u \in [|C|]$) value of the category id belonging to that range. Note that $[n] = \{1, 2, \ldots, n\}$. In order to ensure that all subgraphs are edge-disjoint, we reassign each edge in the current set of edge-weighted subgraphs to a new subgraph G_{r_a,r_b}^{EW} that contains only those edges $c_i \to c_j$, where $c_i \in r_a$ and $c_j \in r_b$. More formally, G_{r_a,r_b}^{EW} is defined as follows.

Definition 5 `(edge-weighted range category graph)`
Given two ranges $r_a, r_b \in \mathcal{R}$ (where possibly $a = b$), the edge-weighted range category graph G_{r_a,r_b}^{EW} is the new edge-weighted category graph containing precisely those weighted edges $c_i \to^{e_k} c_j$ in G^{EW} for which $c_i \in r_a$ and $c_j \in r_b$.

Assuming R ranges, this results in $R(R+1)/2$ edge-disjoint edge-weighted range category graphs. The union of these graphs is a new edge-weighted category graph $G^{EW} = \bigcup_{r_a,r_b \in \mathcal{R}} G_{r_a,r_b}^{EW}(V, E)$, which corresponds to the Wikipedia category network; i.e., our partitioning operations do not lose any information.

4.2 Phase II: Filtering

In the second phase, we introduce a threshold parameter $t \in \mathbb{N}$ that will be used to obtain the filtered category graph G_t^{EW}. To do this, for each $G_{r_a,r_b}^{EW} \subset G^{EW}$, all edges with weight less than t are removed. Hence, each $G_{r_a,r_b}^{EW_t}$ is converted to its corresponding t-filtered category graph $G_{r_a,r_b,t}^{EW}$. It is easy to see that $G_t^{EW} = \bigcup_{r_a,r_b \in \mathcal{R}} G_{r_a,r_b,t}^{EW}(V, E)$. Note that during this phase all isolated categories are removed.

4.3 Phase III: Merging

In the third phase, we first identify the connected components within each $G^{EW}_{r_a,r_b} \subset G^{EW}$ using Breadth First Search (BFS). Each connected component corrresponds to a category cluster $\mathcal{C}(G^{EW}_{r_a,r_b}, t)$, by Definition 4. Merging all connected components of all these subgraphs of G^{EW}_t by combining components that share at least one category, we obtain the complete set of category clusters for threshold t as the connected components of G^{EW}_t. Finally, the obtained category clusters are merged by connecting pairs of clusters that share at least one category into a singe category graph. This process is repeated until all pairs of category clusters are disjoint.

5 Experiments

5.1 Setup

We used the English Wikipedia category network for evaluating the performance of the proposed framework. We studied three years: 2010, 2011, and 2012. Each year was studied separately as an individual dataset. The data is freely available online [2].

For each year we used the same number of partitions of the initial page-category graph, i.e., $N = 2,000$. During this process we eliminated all isolated pages and then transformed each page-caterogy graph to its equivalent edge-weighted category graph by assigning the edge weights accordingly and eliminating all isolated categories. Next, we eliminated duplicate instances of category pairs within different partitions by spliting the categories into ranges (as indicated by Phase I of the framework). We used 70 ranges, i.e., $R = 70$, resulting in a total of $2,485$ edge-disjoint edge-weighted range category graphs.

In addition, we studied different values for the t threshold, ranging from 2 to 4096. Note that all feeble edges (having weight equal to 1) were removed from the network as required by the framework.

The framework was implemented in Java on an Intel i5 processor. The execusion time depends critically on the number of edges and the size of the clusters. For example, for $t = 2$ it took over a week, while for $t = 4096$ it took less than a minute to perform all the computations.

5.2 Results

Our experimental findings on the three English Wikipedia category networks are presented next. We present the structural properties of the networks and investigate the structural behaviour of the clusters with respect to the threshold t.

[2] http://dumps.wikimedia.org/index.html

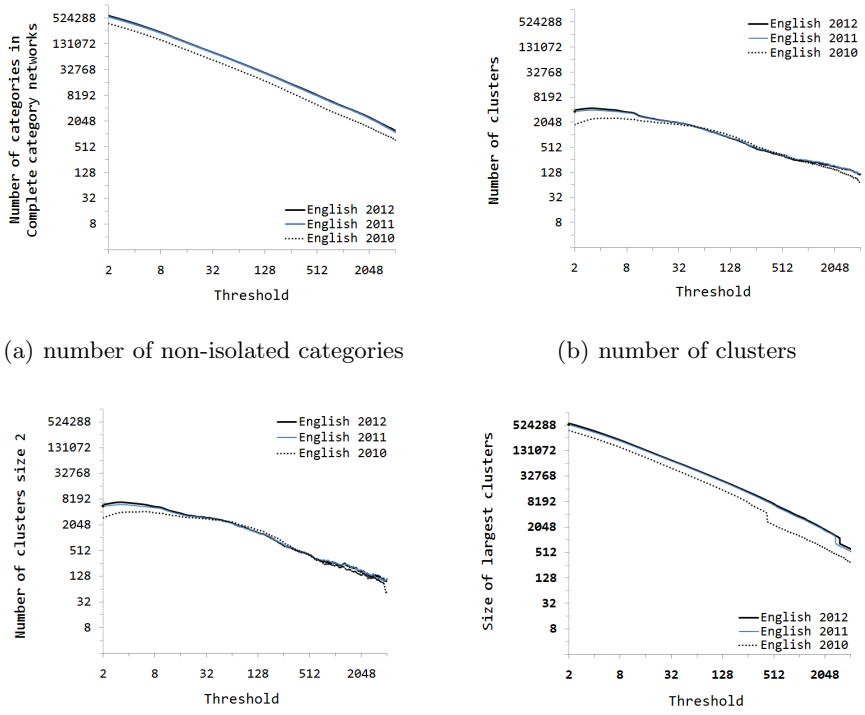
Structural Properties The structural properties of the three Wikipedia category networks are summarised in Table 1. It can be observed that, from 2010 to 2012, the number of pages and categories increased by around 40% and 50%, respectively. However, it is interesting to note that, although the number of isolated pages increased by around 60%, the number of isolated categories was almost unchanged. A possible explanation for this is that, when new categories are added to the network, they are likely to be linked to existing pages as well as new pages. They will therefore be related to existing categories. We also note that the number of page-category links (i.e. edges in the page-category graph) increased by around 50%. A consequence of this, which can be checked using Table 1, is that the average degrees, both pages per category and categories per page, were substantially unchanged.

Table 1: Structural Properties of English Wikipedia Category Link Networks

Network Properties	English 2010	English 2011	English 2012
Number of pages	8,989,264	12,182,689	12,453,596
Number of categories	567,939	801,902	858,869
Number of page-category links	39,484,287	56,969,309	60,386,600
Number of isolated pages	1,083,655	1,735,857	1,755,160
Number of isolated categories	7,443	7,858	7,375
% Isolated pages	12.05%	14.25%	*14.09%*
% Isolated categories	1.31%	0.98%	*0.86%*

Structural Behaviour of Category Clusters We studied how the category clusters depend on the weight threshold t. Specifically, we studied all values of t from 2 to 4096. Some of our most important findings are shown in Figure 3 where we can see that four log-log plots follow a power law. The charts show the number of categories (excluding isolated categories) in the complete category networks in (a), the number of the category clusters in (b), the number of clusters of size 2 in (c), and the size of the largest clusters in (d). A very significant finding here is that all these four log-log plots appear to exhibit power-law behaviour with respect to t. It also seems that the power law exponent is not significantly changing over the time period studied in this paper.

In addition, we note in Figure 3(d) that there is a threshold value for each of the three datasets where the size of the largest category cluster drops sharply. This suggests that each large category (for each of the three years) has diffused into smaller categories or a subcategories. Taking a closer look at these diffusion points, we observed that, in all three cases, the largest cluster was split into two large subclusters. Hence, we plot those diffusion points and display them individually for the 2010 (Figure 4) and 2012 (Figure 5) networks. Due to space limitations, we omit year 2011. In both figures we can see the significant diffusion



(a) number of non-isolated categories

(b) number of clusters

(c) number of smallest clusters

(d) size of largest cluster

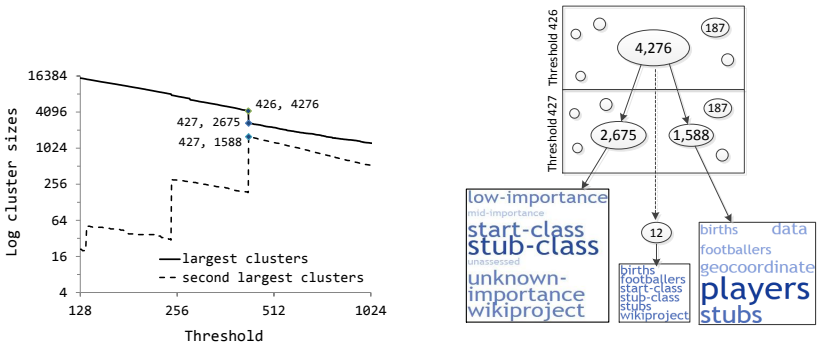
Fig. 3: Log-log plots of (a) the number of non-isolated categories, (b) the number of category clusters, (c) the number of clusters of size 2, and (d) the size of the largest cluster, for different weight threshold values for the English Wikipedia Category Network 2010 - 2012.

points (t threshold values) and the corresponding sizes of the first and second largest category clusters.

Semantics of the Cluster Diffusion We studied the semtantics of the category cluster diffusion. Specifically, we compared the categories that appeared in the original large cluster and then those that appeared in the two largest clusters right after the diffusion. Almost all categories were preserved before and after the diffusion, but were split between the two clusters—so very few categories diffused into smaller components.

In addition, we note that, after the diffusion, a small fraction of the categories present in the initial cluster were not part of any of the new diffused clusters. In the case of year 2010, these were twelve categories missing, for 2011 there were none, while for 2012 there was only one.

Hence, based on the previous observation, we investigated whether there exists any semantic connection or relation between the categories within the two diffused clusters. Specifically, we observed that the frequent categories in the two clusters were substantially different. One cluster would typically contain more general category types, such as "start-class", "stub-class", "people", and "articles", while the second cluster would contain more specific category types, such as "players", "american articles", and "footballers". Some examples of the dominant category titles can be seen in Figure 4(b) for the 2010 network and in Figure 5(b) for the 2012 network.



(a) sizes of largest and second largest clusters

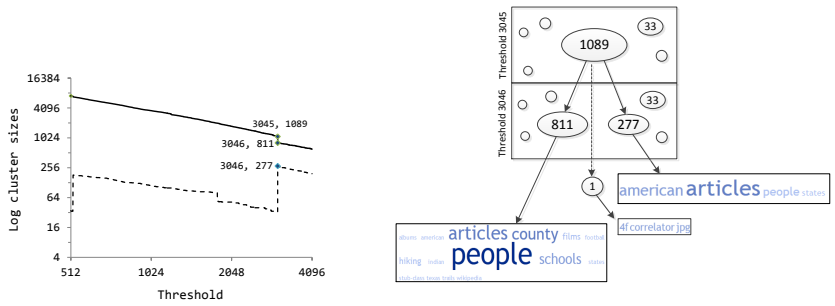
(b) diffused category titles

Fig. 4: (a) English Wikipedia Category 2010 Log-log plots of the largest and second largest cluster sizes and (b) examples of the category titles of the diffused clusters.

6 Summary and Conclusions

In this paper we presented a framework for manipulating a large Wikipedia page-category network. The proposed framework was used to analyze the structure of the network. We obtained, in the Wikipedia category network, global category clusters in form of well-connected components.

In our experiments, we demonstrated the applicability of the proposed framework to several instances of the English Wikipedia category network and observed that, over the years 2010 to 2012, the number of pages, categories, page-category links and isolated pages all increased by 40-60%, but the number of isolated categories was fairly constant. The most significant finding was that the number of non-isolated categories, the number of clusters, the number of clusters of size two, and the size of the largest cluster, and the size of the largest all



(a) sizes of largest-second largest clusters

(b) diffused category titles

Fig. 5: (a) English Wikipedia Category 2012-Log-log plots of the largest and second largest cluster sizes and (b) examples of the category titles of the diffused clusters.

appear to follow power laws with respect to the threshold t. This behaviour is observed for each of the three years of the English Wikipedia category network studied in the paper. Furthermore, for each network we found the value of the threshold t for which increasing the threshold caused the largest cluster to diffuse into two smaller category clusters of significant size. We also observed that this diffusion is typically the result of a "giant" cluster splitting into smaller sub-clusters.

Future work includes the study of our framework on other languages of the Wikipedia category networks. Based on our current investigation of the Wikipedia category graphs, several other languages appear to show a similar cluster structure. In addition, other possible graph clustering techniques are being considered, in particular, the k-core of the category graph and how it relates to the components of the t-filtered category graph.

References

1. I. Beschastnikh, T. Kriplean, and D. W. McDonald. Wikipedian self-governance in action: Motivating the policy lens, 2008.
2. L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph, 2006.
3. S. Goel, J. M. Hofman, and M. I. Sirer. *Who Does What on the Web: A Large-Scale Study of Browsing Behavior*. 2012.
4. M. Hu, E.-P. Lim, and R. Krishnan. *Predicting Outcome for Collaborative Featured Article Nomination in Wikipedia*. 2009.
5. Y. Jiali, J. Liping, Y. Jian, H. Houkuan, and Z. Ying. Document topic extraction based on wikipedia category. In *Computational Sciences and Optimization (CSO), Fourth International Joint Conference*, pages 852–856, 2011.

6. D. Jurgens and T.-C. Lu. *Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia*. 2012.
7. J. Kamps and M. Koolen. Is wikipedia link structure different?, 2009.
8. A. Kittur, E. Chi, and B. Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1509–1512. ACM, 2009.
9. D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. *When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages*. 2011.
10. J. Leskovec, D. Huttenlocher, and J. Kleinberg. *Governance in Social Media: A Case Study of the Wikipedia Promotion Process*. 2010.
11. S. A. Macskassy. *On the Study of Social Interactions in Twitter*. 2012.
12. P. Massa. Social networks of wikipedia, 2011.
13. A. Sadilek, H. Kautz, and V. Silenzio. *Modeling Spread of Disease from Social Interactions*. 2012.
14. S. E. Schaeffer. Graph clustering, 2007.
15. P. Schonhofen. Identifying document topics using the wikipedia category network, 2006.
16. J. Scott. *Social Network Analysis A Handbook*. SAGE Publications, London, 2011.
17. K. Suchecki, A. A. A. Salah, C. Gao, and A. Scharnhorst. Evolution of wikipedia's category structure. *Advances in Complex Systems*, 15, 2012.
18. Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. *The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions*. 2012.
19. M. Wattenhofer, R. Wattenhofer, and Z. Zhu. *The YouTube Social Network*. 2012.
20. T. Zesch and I. Gurevych. Analysis of the wikipedia category graph for NLP applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.