Community Structures in Wikipedia

Ryan Burton
ryb@umich.edu
School of Information
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

Wikipedia, seemingly more than other Web sites, has a rich link structure that seems ripe for analysis. The communities that might exist in Wikipedia may be evident in the network's structure. This paper investigates the network structure of Wikipedia, and explores the degree to which the community structures map to the tags that del.icio.us users annotate pages with. The extent to which Wikipedia's communities follow the cluster hypothesis is explored, along with potential ways to use community structures in retrieval.

1. MOTIVATION

Traditional clustering methods have a long history in information retrieval (IR), and have been the subject of frequent study. Typical clustering methods have large time complexity, making them more suitable for smaller collections. In addition, clustering methods tend to ignore hyperlinks between documents[1].

Network theory has seen a relatively recent growth in interest, and one topic of study in the field is that of community structures. The essence of a "community" is that entities in a community are more similar to each other than an entity is to another outside of its community. This has obvious parallels to clustering which, in IR, relates to a hypothesis of relevance known as the *cluster hypothesis*.

There is a wide variety of algorithms for finding communities in networks—each having a different discipline of origin and assumptions to go with them. These algorithms ignore everything but the graph structure that the network presents, and we aim to analyse its effectiveness in doing so.

2. COMMUNITY DETECTION METHODS

The notion of a community is intuitively simple—it is a structure that segments a network such that there is a higher density of edges within each group than between them [3]. Although this is the case however, formalising the notion of a community is not as trivial, and many definitions exist – none of which are considered standard [1]. In fact, we will explore two separate notions of communities in this paper. To further complicate matters, "community detection" may also be referred to as "clustering" in computer science and sociology [1], and "communities" may also be called "modules" in the literature. This paper will adopt the former terms in each case.

Because of the large number of community detection algorithms that exist, only two methods were tested – CFinder[2], based on clique percolation [5], and an implementation of the "Fast Modularity" algorithm[3] described by Clauset, et al. [1]. An advantage of both these methods is that they *find* community structures – both the groups that each vertex belongs to, as well as the number of groups. This is different from algorithms such as k-means, which requires the number of communities to be specified in advance.

2.1 CFinder

CFinder is an implementation of the clique percolation method described by Palla, et al. [5]. The essence of the algorithm involves defining a *k-clique*, which is justified by the hypothesis that groups within a community are dense, and thus likely to form cliques. A k-clique is a complete subgraph (all possible edges are present), which moves or "rolls" on a graph; if the clique is trapped and cannot proceed, the inter-community edges form a bottleneck, and it is possible to consider the extent of its travel as a community.

The algorithm defines a number of concepts – one is the adjacency of k-cliques. Two cliques are adjacent if they share $k-1$ vertices. The "roll" occurs when the k-clique rotates about the $k-1$ vertices. The *k-clique chain* is the union of adjacent k-cliques (which can be thought of as the path that it rolled along). A k-clique community is the union of the k-clique and the resulting k-clique chain such that the community is a connected subgraph [3].

A positive aspect of CFinder is that it works on both directed and undirected networks. This makes it suitable for Web documents, which contain directed hyperlinks. The main disadvantage however, is that it can be very slow; its time complexity is $O(\exp(n))$ [2]. The result is that running the complete algorithm did not finish in time for this paper.

2.2 Fast Modularity

"Fast Modularity", as it is called on its home page, is a freely available implementation of the community structure

[1]If we assume that hyperlinks imply similarity, then a sophisticated similarity/distance metric could take this into account as well.

[2]http://cfinder.org/

[3]http://cs.unm.edu/~aaron/research/fastmodularity.htm

inference algorithm described by Clauset, et al. [1]. It is an instance of hierarchical agglomeration – each node is first assumed to be its own community, and iteratively merges communities based on a *modularity* measure until one community remains to form a dendrogram. It records the modularity at each step to help determine the cut point for the dendrogram – the step at which modularity is maximised would make a suitable cut point.

Modularity is a measure of the probability of within-community edges existing due to chance. The formula for modularity is as follows (where v and w are vertices):

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

The choice of data structures in the Fast Modularity program makes it fast enough for Web network data on the order of 10,000 vertices; the running time is $O(md \log n)$, where n and m are the number of vertices and edges respectively, and d is the depth of the dendrogram of the actual hierarchy. The authors note that many real-world networks are rather sparse and hierarchical ($m \sim n$, $d \sim \log n$), which would make the running time $O(n \log^2 n)$, or "essentially linear".

The Fast Modularity algorithm expects undirected graphs which makes this less suitable for the Web than clique percolation. The methods used to induce something closer to an undirected network will be described later in the paper.

3. WIKIPEDIA'S GRAPH STRUCTURE

Wikipedia[4] is a free encyclopedia with a liberal licensing scheme that any user can edit. The articles within it form a rich network structure that, especially because of its size, makes it relatively interesting to study. Everything2[5] is a similar site in that it encourages users to create and edit articles, with some fundamental differences. One major difference is that Wikipedia is based on several policies – one such policy is its linking policy[6] which encourages users to strike a balance between over-linking and under-linking. In comparison, Everything2 is more liberal and its users treat the site as such, making links at times without as much regard for whether this link at the other end exists (because the article, or "node" has been written, or whether the title is the same as the anchor text, etc.), or whether the link is truly relevant to the subject. We would expect therefore that links in Wikipedia imply a strong semantic relationship.

The English Wikipedia[7] has within it 3,268,130 articles at the time of this writing. A somewhat more manageable subset is the Wiki10+ Dataset [8], which contains 20,764 articles and Delicious[9] tag annotations for each article. Some network statistics that describe the data set are as follows:

- Nodes: 20764
- Edges: 772904
- Density: 0.0017928
- Mean Clustering Coefficient: 0.1664698
- Average distance among via reachable pairs: 3.49912
- Number of unreachable pairs: 16.6 million
- Diameter: 9
- Average in-degree: 4
- Average out-degree: 19

The network has a low average distance and high clustering coefficient, which are properties of small-world networks. In comparison, a random graph has a mean clustering coefficient of 0.006686. Plots of the in-degree and out-degree distributions are included in Figures 1 and 2 respectively. The in-links seem to follow a power law distribution, whereas the out-links are based on a Poisson process. We would indeed expect a select few pages to be referenced the most, and most pages to have a similar number of references.

It should be noted that the original data set has an article that is referenced by every other page in the data set. This was removed for the community detection task.

3.1 Removing Directed Edges from the Network

Because the Fast Modularity algorithm expects undirected edges, it was necessary to remove directed edges. Exploration of the motifs present in the data showed that reciprocal links were very common – it is possible to consider these undirected edges. As such, all other edges were removed, which resulted in a network of 18738 nodes and 118026 edges. Using co-citation to induce an undirected network may likely be a better method, but it was not feasible to perform this on such a large network due to time and memory constraints.

4. COMMUNITIES IN WIKIPEDIA

The Fast Modularity algorithm by Clauset, et al. found 205 communities in the reduced data set of undirected edges. The smallest community has two articles, while the largest has 4241 – a rather large variance. Communities have a mean of 91.4 articles.

The default behaviour for CFinder is to automatically try different values of k. These can actually result in vastly different community structures. The expectation may be that the operator will explore the different candidate community structures, but it was not possible to do so with the CFinder program[10]. Looking at the results, it seems that CFinder can find as many as 1005 communities at $k = 3$ (one is an approximate community, with 11076 vertices; most of the other 1004 communities are small).

One of the goals of this paper was to use del.icio.us tags as indicators of similarity; it seemed possible to use the Jaccard coefficient on the set of tags for each article:

$$J(T_{d1}, T_{d2}) = \frac{T_{d1} \cap T_{d2}}{T_{d1} \cup T_{d2}}$$

However, this turned out to be less useful than expected because of the lack of structure that a folksonomy has. As an

[4]http://www.wikipedia.org/
[5]http://everything2.com/
[6]http://en.wikipedia.org/wiki/Wikipedia:OVERLINK
[7]http://en.wikipedia.org/
[8]http://nlp.uned.es/social-tagging/wiki10+/
[9]http://www.delicious.com/

[10]The visual interface runs out of memory, and the text files that the command line program generated had to be examined.

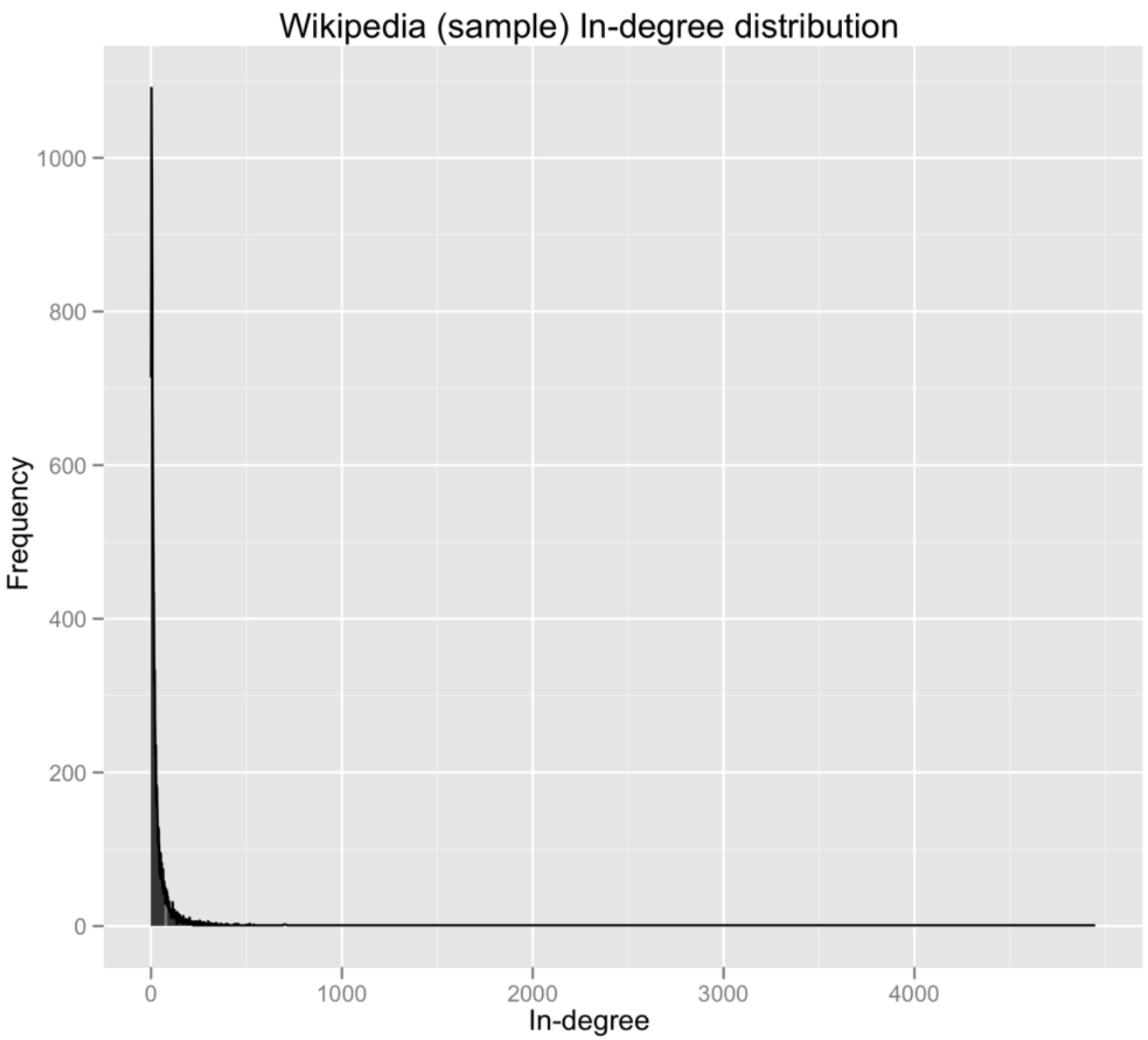

Figure 1: In-degree distribution of a Wikipedia network

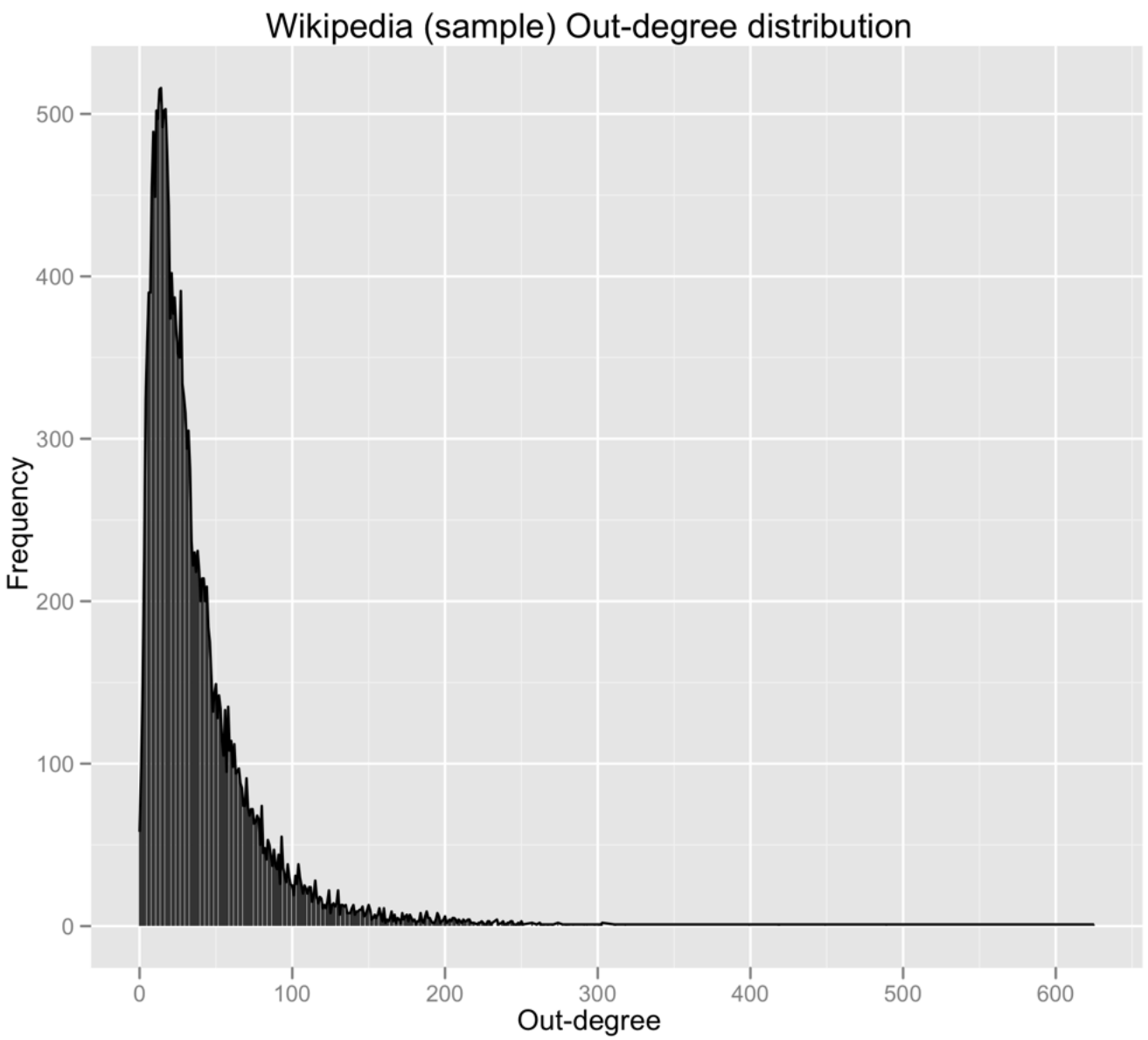

Figure 2: Out-degree distribution of a Wikipedia network

example, the Fast Modularity algorithm found a community of language lists (shown are article titles):

> ["Veni, vidi, vici", "List of Latin words with English derivatives", "Q.E.D.", "List of Latin phrases", "List of Greek phrases", "List of German expressions in English", "List of Latin phrases (full)", "Greek and Latin roots in English", "List of Latin abbreviations", "List of English words of Yiddish origin"]

However, the only intersecting tags of this group are "reference" and "wikipedia". There are 148 tags total for these pages, with some being non-obvious or potentially relevant only to the person who tagged it (examples are "bloggable" and "cherryblossom"). Even some tags used like "language" or "linguistics" are not common.

There are nevertheless good examples of tagging implying similarity, but these apply to the smaller communities:

> Community: ["Internet fax", "T.38"]
>
> Common tags: ["fax", "foip", "lunchfax", "reference", "voip", "work"]
>
> All tags: ["client_scp_fax", "computer", "fax", "faxing", "foip", "internet", "kmbsit", "lunchfax", "reference", "software", "t38", "telecommunication", "telefonia", "toip", "voice", "voip", "wiki", "work"]

Another is:

> Communities: ["Frederick Herzberg", "Two factor theory"]
>
> Common tags: ["herzberg", "leadership", "management", "motivation", "psychology", "theory", "work"]
>
> All tags: ["(2)behaviouralstudies", "altexor", "b830", "business", "business_start-ups", "communitybuilding", "culture", "design", "dissatisfaction", "effectiveness", "employment", "eppp", "factor", "finance", "frederick", "frederick_herzberg", "herzberg", "humanresources", "hygiene", "info5570", "interesting", "job", "leadership", "life", "lifehacks", "management", "mangement", "marketing", "maslow", "money", "motivation", "needs", "pmp", "psychology", "pub_ad", "pygmalion", "reading", "reference", "satisfaction", "theory", "twofactortheory", "wikipedia", "work"]

Do note in the last example however the more idiosyncratic tags like "(2)behaviouralstudies" and "info5570". Even "interesting" is troublesome to account for.

Larger communities tend to have no common tags—this may be due to the potential for unrelated articles to creep in:

> ["Mothman", "Coca-Cola", "Jackalope", "Kitsune", "Carnivorous plant", "Vegetable Lamb of Tartary", "Mongolian Death Worm", "Cryptid", "Japanese mythology", "Qi Xi", "List of legendary creatures", "Logo" ,"Cicada", "The Gods Must Be Crazy", "Symbol grounding", " Gigantopithecus", "Montauk Monster", "Venus Flytrap", "Man-eating tree", "Qilin", "Thylacine", "Tanabata", "The Snow Queen", "Yotsuya Kaidan", "Journey to the West", "Xuanzang", "Yuki-onna", "List of cryptids", "Bigfoot", "Globster", "Loch Ness Monster", "Raymond Loewy", "Chinese mythology", "Symbol", "Coca-Cola formula", "List of legendary creatures from Japan", "Kappa (folklore)", "Slow Down (unidentified sound)", "Bloop", "Romance of the Three Kingdoms", "Alpaca", "Japanese folklore", "The Hum", "Kami", "Yeti", "Four Great Classical Novels", "Tulpa", "Dream of the Red Chamber", "Jersey Devil", "OK Soda", "Water Margin", "Jiang Shi", "Kraken", "OpenCola", "Animal Crossing", "Tanuki", "Buddhas of Bamyan", "Symbolism", "Okapi", "Tengu", "Chupacabra", "Tasmanian Devil"]

In this group of articles which are seemingly about folklore and legends, "Coca-Cola", "Coca-Cola formula", "Open Cola" and "OK Soda" are present, though they should perhaps be separated out into their own community. "Animal Crossing" is a video game that is arguably not a part of this group, though there is an interesting (but rather weak) connection.

5. APPLICATIONS OF COMMUNITY DETECTION TO INFORMATION RETRIEVAL

A point of inquiry for our purposes is whether community structure is a good indicator of similarity between documents. The *cluster hypothesis* is a principle in information retrieval that posits that relevant documents are likely to be similar to each other [4]. If it is the case that two articles belonging to the same community implies similarity, then the cluster hypothesis would apply to network communities as well, and we can use it as a surrogate for traditional clusters when, for instance, dealing with large amounts of data.

In order to examine the correlation between community presence and similarity, it is possible to observe the distribution of similarities within and among groups [6]. The similarity metric used for this task was *cosine similarity*, defined as:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Using this, the distribution of similarities within communities was plotted, as seen in Figures 3 and 4. Because of time constraints, only a sample of the potential similarities were computed and plotted[11].

5.1 Potential for "clustering" search results

One use of clustering is that of grouping search results and presenting these to the user of an IR system. An example of one such search engine is Clusty[12]. This has the advantage of grouping different concepts together, so that the polysemy problem is mitigated – grouping Jaguar the operating system differently from Jaguar the make of automobile.

[11] In the case of the stemmed documents, the stemmer appears to somehow have a memory leak – as many similarities were computed until memory was exhausted.

[12] http://www.clusty.com/

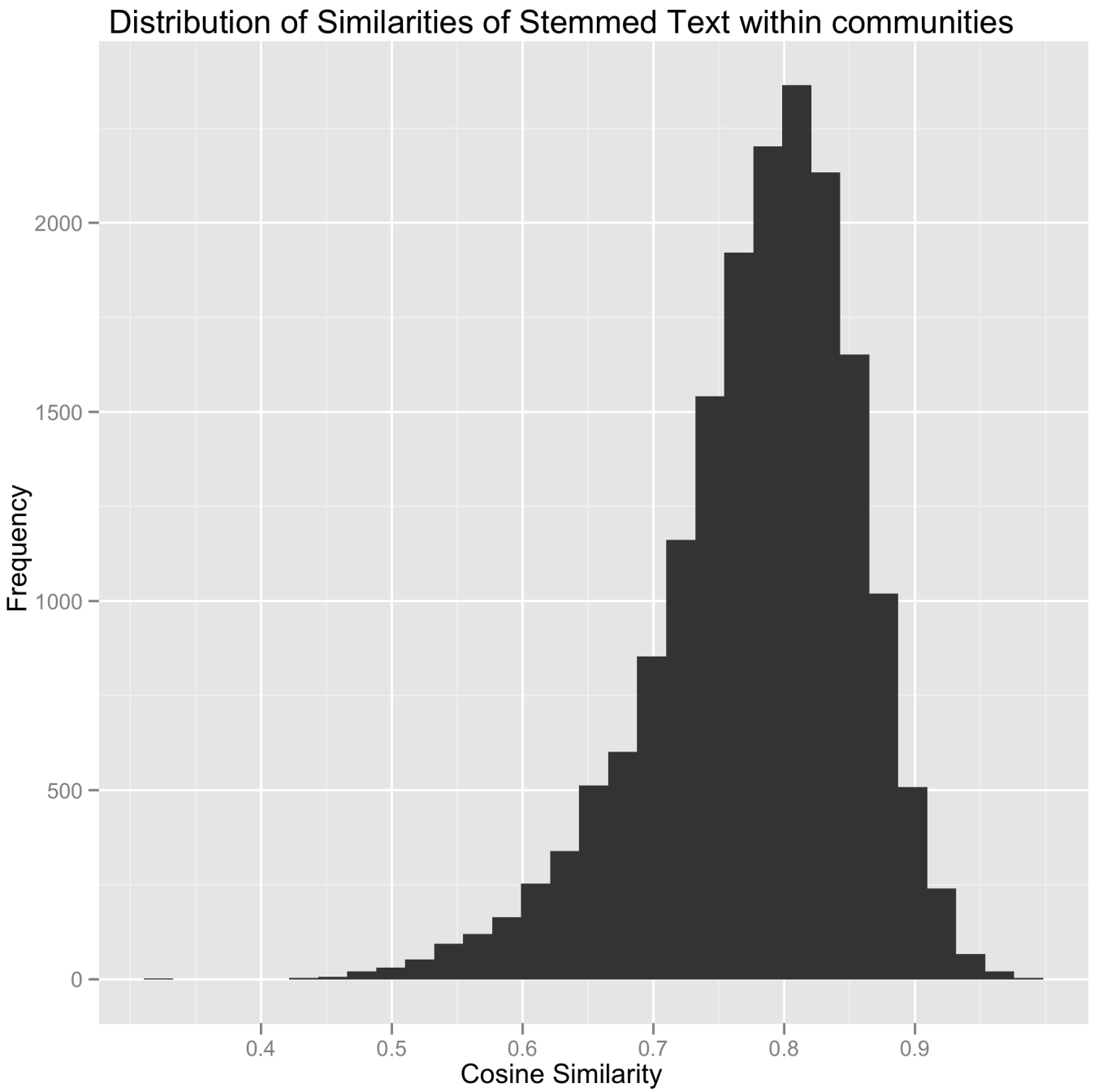

Figure 3: Distribution of similarities within communities of stemmed documents

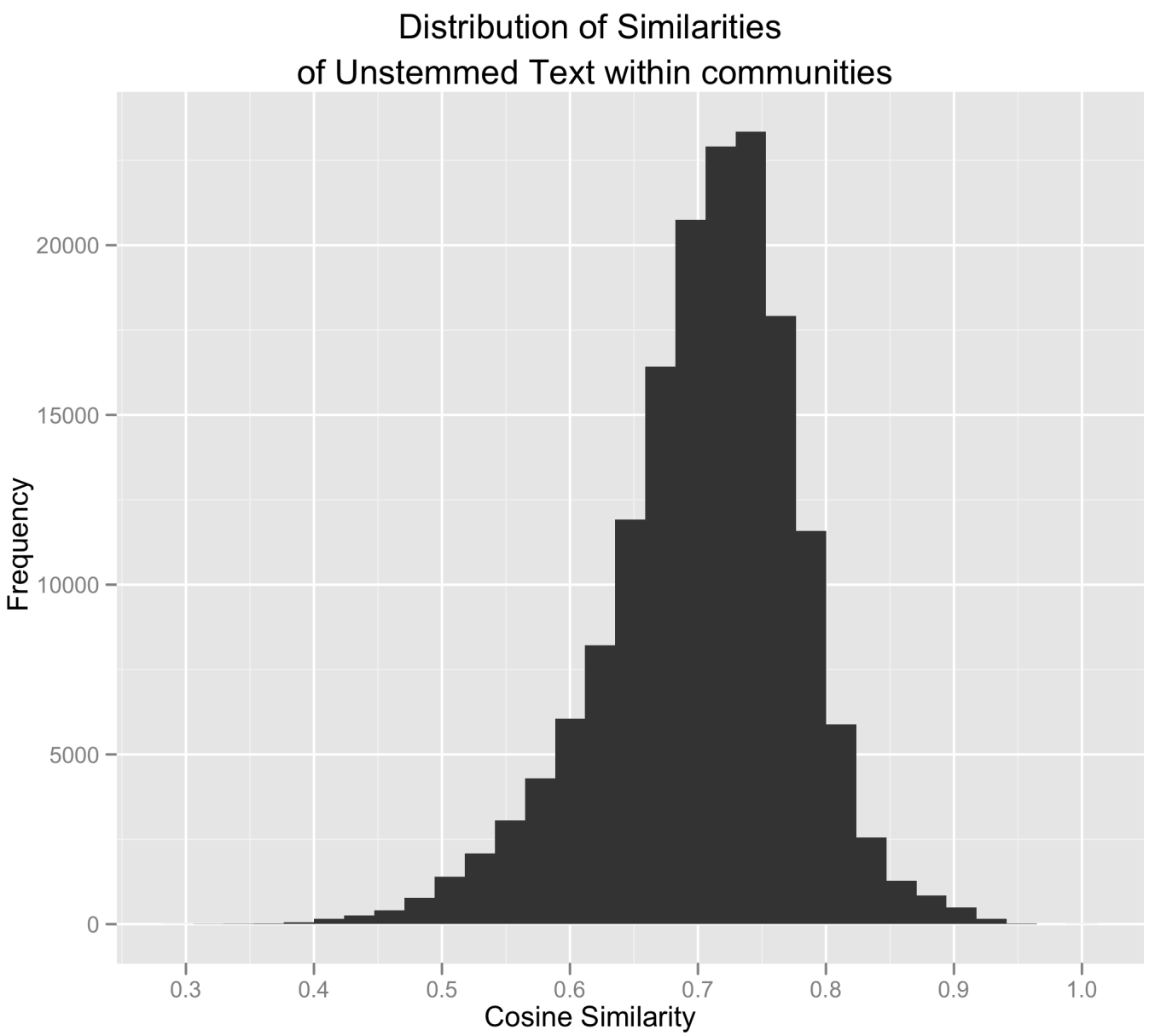

Figure 4: Distribution of similarities within communities of unstemmed documents

To explore this, the article text from the raw HTML data was extracted and indexed with Lucene. A subset of AOL query log data (500 queries) with Wikipedia as the first target was used to query the Lucene index, and the top 1000 results were used to build a graph for each query for 429 graphs – some had no results. At this point, a co-citation matrix was built for each query to induce an undirected graph based on the following calculation (where A is the connectivity matrix):

$$C = A^T A$$

The Fast Modularity algorithm was then used to find communities in these undirected graphs.

The results were quite good, and work especially well for the more vague queries. As an example, one such query is "bob", which results in two clusters:

> ["All Along the Watchtower", "Grateful Dead", "Bob Dylan discography", "Traveling Wilburys", "Theme Time Radio Hour", "Bob Dylan"]
>
> ["Quantum cryptography", "Quantum teleportation", "Bell's theorem", "Man-in-the-middle attack", "Public-key cryptography", "Diffie-Hellman key exchange", "RSA", "Certificate authority", "EPR paradox"]

One is related to music (cf. Bob Dylan), and another is related to cryptography (cf. Alice and Bob[13]). Similarly, the query "neo" gives one community about the Matrix films, and another about art (cf. Neoclassicism), and a query for "the kyoto protocol" (as separate terms and not as a phrase) gives one community about emissions and pollution, and another about network protocols.

6. DISCUSSION AND CONCLUSION

Performing Fast Modularity detection on the full data set yielded some interesting communities, but may not be as representative of the true underlying communities in Wikipedia. The co-citation matrix gives information on whether the article in question is co-cited with another – that is, whether there is an article that links to the two articles. This gives a measure of similarity that may not lead to as much information being lost as might happen by dropping unidirectional arcs. An attempt was made to calculate the co-citation matrix of the full graph, but memory constraints posed an issue. Presumably building a matrix of the full graph would require less than 2 GB of RAM – perhaps more care might be needed in building and calculating the matrices to ensure efficient memory usage.

As might be expected, stemming documents increases the cosine similarity score within communities, and points to the fact that community detection works irrespective of the text in the articles. It would be interesting to look at distributions of more measures of similarity and see which is most similar to presence in a community.

An area of investigation that was planned was determining whether there would have been "similar" documents (according to some other metric) that community detection might have missed. In addition, the plan was made to use the *nearest neighbour* test [6] on the data. The nearest neighbours are most similar to a document – if community detection did its job truly well, we would expect many of these nearest neighbours to be in the same community. Calculating the nearest neighbours was also a very expensive task, and using a subset of the data that finished before this writing did not give particularly meaningful results.

The results of grouping search results were satisfying, but evaluation is indeed difficult without a labelled query corpus. It should be possible to submit the results to a collaborative voting system, or a service like Amazon's Mechanical Turk[14], but just inspecting the results and being uncertain about whether a group truly belongs together seemed to reinforce the point that the person who evaluates the results should be the original source of the information need. As an example, consider the results for the query "anarchy" – it gives two communities:

> ["Anarcho-primitivism", "Anarchism", "Neo-Luddism", "CrimethInc.", "Bob Black", "Derrick Jensen", "John Zerzan", "Anarchist symbolism", "Christian anarchism", "Plutocracy", "List of forms of government", "Corporatocracy", "Anarchy", "Post-left anarchy", "Green anarchism"]
>
> ["Mikhail Bakunin", "Peter Kropotkin", "Mutual Aid: A Factor of Evolution", "Robert Nozick", "Voltairine de Cleyre", "Temporary Autonomous Zone", "Really Really Free Market", "Crypto-anarchism", "Anarchy, State, and Utopia"]

Without some knowledge of the topic, it is not easy to reasonably determine whether these clusters are appropriate or not.

Furthermore, as should be expected, the results are not perfect, and there are often cases when at least one article is classified in the wrong community. For example, the query "monty python" was one that was assessed (as a conjunction of terms and not as a phrase)—it resulted in three communities:

> ["Time Bandits", "Monty Python", "Fluxx", "Douglas Adams", "Eddie Izzard"]
>
> ["Cold open", "Peter Cook", "100 Greatest British Television Programmes", "Beatitudes", "Black comedy", "Python (programming language)", "Monty Python and the Holy Grail"]
>
> ["Guido van Rossum", "Metasyntactic variable", "Stackless Python", "Spam (Monty Python)"]

We can see two obvious cases of misclassification in the second and third communities. Perhaps this task proved troublesome however because of the programming language's heritage of being named after the British comedy troupe.

The Fast Modularity algorithm by Clauset, et al. was again the only of the two that was able to give reasonable results. The graphs are much smaller for search results and are still directed, but CFinder failed to give good results for communities – each instance resulted in one community. To highlight one example, consider the results visualised in Figure 5. We see that the graph seems very dense, which would mean that a k-clique would only percolate throughout one subgraph. The other communities, if they exist, could be expected to be smaller than k, and thus not part of a community.

[13]http://en.wikipedia.org/wiki/Alice_and_Bob/

[14]http://www.mturk.com/

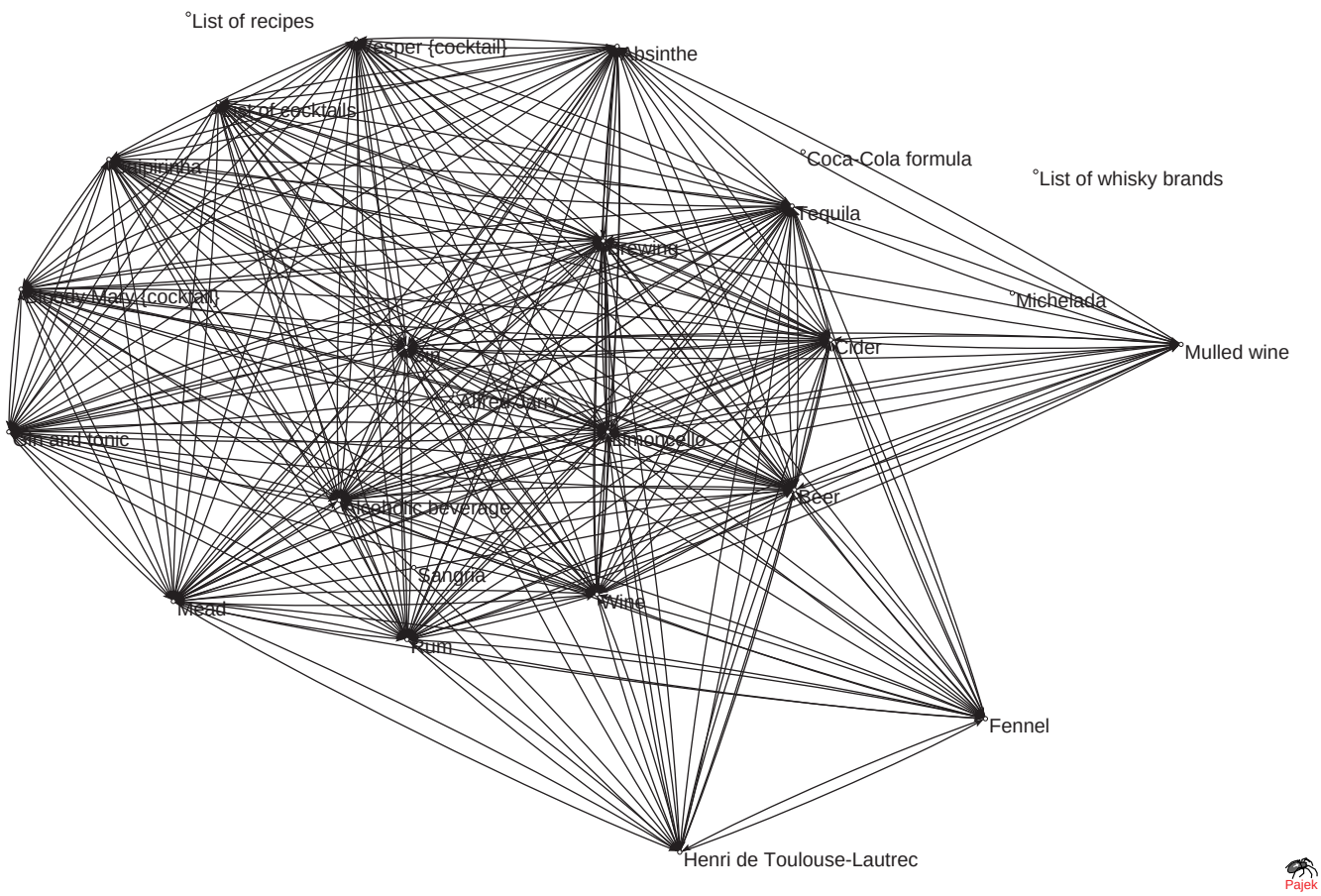

Figure 5: Sample graph of query results

We have seen that the link structure in Wikipedia tells us quite a bit about the similarity of articles. This suggests that the link policy is being upheld, but it would be interesting as well to see whether this is true for other hyperlinked corpora as well. Presence in a community between two documents implies that the documents truly are similar (or related at the very least). Our two community-finding algorithms have interesting properties that might be examined further or exploited in applications that use them – Fast Modularity gives a hierarchical dendrogram (similar to hierarchical agglomerative clustering) and CFinder finds overlapping communities. In our case, we used Fast Modularity to find communities in search results and presented the "best" communities. These would make for interesting topics of future study.

7. REFERENCES

[1] Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks. *Phys. Rev. E 70*, 6 (Dec 2004), 066111.

[2] Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment 2005*, 09 (2005), P09008.

[3] Fortunato, S., and Castellano, C. Community Structure in Graphs. *ArXiv e-prints* (Dec. 2007).

[4] Hearst, M. A. Reexamining the cluster hypothesis: scatter/gather on retrieval results.

[5] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 7043 (06 2005), 814–818.

[6] Voorhees, E. M. The cluster hypothesis revisited.