# The Dynamics of Content Creation in Wikipedia

**Pei-Yu Wang**
anna327@stanford.edu

**Giannis Neokleous**
gneokleo@stanford.edu

**Amir Ghazvinian**
amirg@stanford.edu

## Abstract

*The use of network representations to model and understand data has been applied successfully in a number of diverse areas. In this project, we employ network analysis techniques to understand the dynamics of content creation on the popular online encyclopedia, Wikipedia. We examine several structural properties of a network model corresponding to the revision histories of the content and discussion pages for articles on Wikipedia. We then analyze community structure within the network. We conclude that modeling Wikipedia as a network and analyzing the properties of that network can give useful insights into the way that article content is generated.*

## 1 Introduction

The use of network representations to model and understand data has been applied successfully in a number of diverse areas. In such studies, researchers often generate a network model to represent some data and utilize that model to better understand the properties of that data.

In this project, we employ network analysis techniques to understand the dynamics of content creation on the popular online encyclopedia, Wikipedia [1]. Wikipedia differs from traditional encyclopedias in that the content of the articles can be edited by anyone and is not simply written by a team of professionals or experts. As of 2010, Wikipedia contains over 3.5 million articles, each of which has been written and curated by multiple users. To aid in the process of curation, each article also has a discussion page, where users can discuss how to make the article as useful as possible. Additionally, many of the articles in Wikipedia belong to one or more categories, which give a broad description of the topic of these articles.

By analyzing the structure and properties of Wikipedia as a network, we can better understand which users edit which articles and who talks to whom about the articles. For example, the degree distribution of a network can tell us about the patterns of connectivity within the network.

One property that seems to be common to a lot of networks, and in particular to social networks, is the existence of communities, groups of network nodes within which connections are dense, but outside of which connections are much sparser. The study of community detection in networks can shed light on the organization of networks and their functions. Thus, we also examine community structure in Wikipedia.

Insights into the properties of the network can prove useful in a number of different ways. For example, we can use the results of the analysis to recommend users who would be good candidates to edit or revise a particular article. We could also use the results to suggest groups of users for collaboration on Wikipedia.

Here, in order to better understand content creation on Wikipedia, we perform a number of different types of analysis on data corresponding to the edit and discussion histories of its articles. More specifically:

- We construct network representations of Wikipedia article edits and article discussions to model the data.
- We analyze and interpret the properties of the network, such as the degree distribution, clustering coefficient, and network diameter.
- We use the Clauset-Newman-Moore algorithm [2] to identify communities in Wikipedia. We then analyze the degree to which these communities reflect categories of articles.

## 2 Related Work

Researchers have used the properties of networks to understand dynamics in a number of different domains. For example, Leskovec et al. [3] studied the properties of a product recommendation network in order to analyze a viral marketing campaign based on these recommendations. Additionally, Kwak et al. [4] explored the properties of Twitter as a network. In this paper, the authors analyzed the distribution of follower/following relationships, information diffusion through retweets, and several other characteristics of the network in order to better understand the Twitter service and its power as a forum for sharing information. Both papers follow a number of principles that we attempt to emulate in this study. First, both papers create a network model of the data and do so in a way that corresponds to a real world interpretation. For example, in the work by Kwak and colleagues, one network the authors analyzed represents individuals as nodes and the follower and following relationships between them as edges. Second, both papers analyze properties of the network and interpret the results in such a way as to gain insight into the overall characteristics of the original data. While we use a completely different data set to understand content creation in Wikipedia, the principles and approaches put forth in these papers still applies.

The study of community detection in networks can shed light on the organization of networks and their functions. In community detection, we attempt to identify groups of network nodes such that connections within the group are dense, but connections outside the group are much sparser. There is a large body of work related to identifying communities in networks. One of the most popular community detection algorithms is that of Girvan and Newman [5]. The Girvan-Newman (GN) algorithm uses a divisive approach by removing nodes with the highest betweenness score. The betweenness score is a measure that Girvan and Newman proposed, where the betweenness of an edge is defined as number of shortest paths between all pairs of nodes in the graph that pass through the given edge. This will indicate which edges have many shortest paths passing through them; these edges are then removed from the graph. After each edge is removed from the graph, the GN algorithm recalculates the betweenness for all remaining edges. Because of the complexity of calculating this betweens score, the runtime complexity of the GN algorithm is $O(nm^2)$, so the algorithm is tractable only for a few thousand nodes. Another algorithm, which we use for community detection in this paper, is that proposed by Clauset, Newman and Moore (CNM) [2]. The CNM algorithm uses an agglomerative approach by putting each vertex in its own community and then joining the communities that produce the highest increase in modularity, a metric explained in a subsequent section. The algorithm represents the graph using a multigraph where a vertex represents an entire community where edges between communities are represented by bundles and edges within a community are represented by self-edges on the vertex. This multigraph is represented by an adjacency matrix, which also makes the joining of communities faster since it is just the sum of two rows and two columns. The algorithm has a running time of $O(n\log^2 n)$ and can be applied to networks with millions of nodes.

The work of Andrea Lancichinetti and Santo Fortunato [6] gives a comparative analysis on a variety of algorithms for detecting communities, including the GN algorithm and the CNM algorithm among others. The authors tested the algorithms against the GN benchmark [7], a benchmark community detection test on a graph that consists of 128 nodes, each with expected degree of 16, which are divided into four groups of 32 nodes. They found that modularity based methods such as the CNM algorithm performed particularly well on the benchmark. However, the GN benchmark is not particularly realistic in real life since all nodes have the same degree and all communities are the same size. For this reason, Lancichinetti and Fortunato also performed empirical analysis of community detection on a number of real-world and synthetic graphs, including the famous Zachary's karate club network and, a network of collaboration among physicists, and a network of fictional characters based on the book *Les Miserables*. They found that GN performs

reasonably well under a variety of different network conditions such as in directed graphs, weighted and overlapped networks, and for graphs with different numbers of vertices. CNM also performed well on these tests.

# 3 The Data

In this section, we describe our data set, which consists of edit and discussion histories from Wikipedia. We additionally describe the category hierarchy of Wikipedia, which we use to focus our analysis as well as to evaluate communities we detect in our network representation of the data (Section 5).

## 3.1 Wikipedia Data

For this analysis, we use two data sets corresponding to Wikipedia revision histories. The first data set, which we call edit data, consists of revisions to the content of articles in Wikipedia. The second data set, which we call talk data, consists of revisions to the discussion pages for articles and therefore indicates who is talking about the articles.

Each entry in the data set corresponds to a single revision to either the content or the discussion page of an article. Figure 1 shows a sample entry in the data set and demonstrates the format of the revisions. In this analysis, we are interested only in the title of the article being revised, the time of the revision, the user making the change and the category of the article. We ignore the remaining information contained in the entry, such as other articles that are linked to by the article being revised or the comments associated with the revision.

Because the total number of entries in the data set is exceptionally large, we limit our analysis to revisions made during the six month period from July to December 2007. We also exclude all revisions for which the article does not have any category, since we use the categories as a metric for evaluation of communities in the data set. For the edit data set, we further limited our data set to only include entries referring to articles in three specific categories: science, arts, and places. Section 3.2 gives a detailed description of the Wikipedia category hierarchy and how we used the hierarchy to filter out articles that were not included in one of these categories. Table 1 describes the edit and talk data sets that we use and gives some summary statistics of the data.

```
REVISION 4781981 72390319 Steven_Strogatz 2006-08-28T14:11:16Z SmackBot
433328
CATEGORY American_mathematicians
MAIN Boston_University MIT Harvard_University Cornell_University
OTHER De:Steven_Strogatz Es:Steven_Strogatz
EXTERNAL http://www.edge.org/3rd_culture/bios/strogatz.html
TEMPLATE Cite_book Cite_book Cite_journal
COMMENT ISBN formatting &/or general fixes using [[WP:AWB|AWB]]
MINOR 1
TEXTDATA 229
```

**Figure 1.** This figure shows a sample entry from the Wikipedia data set. Each row consists of a title written in capital letters, followed by data for that row, which is delimited by spaces. The row called revision indicates the title of the article, the date of the revision, and the user who revised the entry. In this example, the article title is "Steven_Strogatz", the timestamp for the revision is "2006-08-28T14:11:16z", and user "SmackBot" made the revision. The row called category indicates the category of the article, which is "American_mathematicians" in this example. For the purposes of this analysis, we are only interested in the two rows described here.

| | **Edit** | **Talk** |
|---|---|---|
| **Dates** | July - Dec 2007 | July - Dec 2007 |
| **# Entries** | 751,376 | 161,521 |
| **# Articles** | 34,066 | 12,406 |
| **# Users** | 195,953 | 26,125 |

**Table 1.** This table describes our edit and talk data sets, which correspond to revision histories of articles and their discussion pages. Each entry in the data set is of the format described in Figure 1.

## 3.2 Wikipedia Category Hierarchy

Categories in Wikipedia enable articles to be tagged and added to automatic listings. Thus, categories help structure Wikipedia by grouping together pages on similar subjects. Each article may be tagged with zero or more categories.

Wikipedia facilitates the use of categories to group related articles through a user generated category hierarchy. This hierarchy can be thought of as a series of overlapping trees. Any category may have several subcategories and it is possible for these subcategories to have more than one parent category. For example, a category *A* can have subcategories *B, C*, and *D*; however, category *B* may have another parent *E* in addition to having *A* as a parent. The tree hierarchy can also contain cycles. For example,

category D may have a subcategory F whose subcategory is category A.

In total, Wikipedia contains more than 450,000 categories. We obtained a relational database containing all these categories as well as every parent-child relationship between pairs of Wikipedia categories [8]. We use these categories to focus the set of articles we include in our analysis of the edit revision history as described below.

When filtering out the articles to keep for the creation of our edit network, we chose three categories as described above: science, art, and places. We chose these categories because we believed that they would give a wide range of different articles contributed by many users with very different expertise. To filter the articles based on these categories, we kept in our edit data set all articles that were tagged either with the category directly or with one of the category's children or grandchildren in the hierarchy. We did not go any further down the category tree than the grandchildren of the original category. We restricted to grandchildren because the size of the trees increases dramatically at each level and we therefore could not include any more articles in the graph.

## 4 Network Properties

In this section, we describe our model of Wikipedia as a network of connected articles as well as the characteristics and properties of that network. We discuss how each of these properties helps us better understand the structure of the network and, by extension, Wikipedia.

### 4.1 Modeling Wikipedia as a Network

We model Wikipedia as a network in which each node represents an article. When using the edit data as described in Section 3, each edge in the network indicates that an individual edited the content of both articles connected by the edge. When using the talk data, which consists of revisions made to the discussion page for an article, we connect two articles when the same individual has edited the discussion page for both articles. The edit network therefore reflects connections among the content revisions in Wikipedia and the talk network shows connections among article discussions.

To create each network, we first processed the raw entries in our data set and filtered out all articles without categories listed. We did this because we wanted to be able to evaluate the extent to which Wikipedia authors edit articles with similar subjects, and we use the category hierarchy of Wikipedia to do so. After filtering out articles without categories, we further processed the remaining revisions to create a file in which each line gives the names or ids of two nodes that are connected by an edge. Once we had a file in this format, we could use it to generate a graph and analyze the properties of the graph.

Table 2 describes the networks we generated to represent our edit and talk data sets. As shown in the table, the edit network is much larger than the talk network, even given that the edit network only contains articles in the categories science, art, and places. Additionally, nodes in the edit network typically have a much higher degree, averaging about 1,173 edges per node as opposed to 36.5 edges per node in the talk network.

| | **Edit** | **Talk** |
|---|---|---|
| **Nodes** | 34,066 | 12,406 |
| **Edges** | 39,960,455 | 4,533,222 |
| **Avg degree** | 1,173.03 | 36.5 |

**Table 2.** This table describes our edit and talk network models. For each network, the table gives the number of nodes, number edges, and the average degree of each node.

### 4.2 Degree Distribution

Given the network model described above, we first computed the degree distributions for both the edit and talk networks that we constructed. Degree distribution shows for a given degree k, the number of nodes in the network that have exactly k edges. Figure 2 shows plots of the degree distribution as a complementary cumulative distribution function for both the edit and the talk networks.

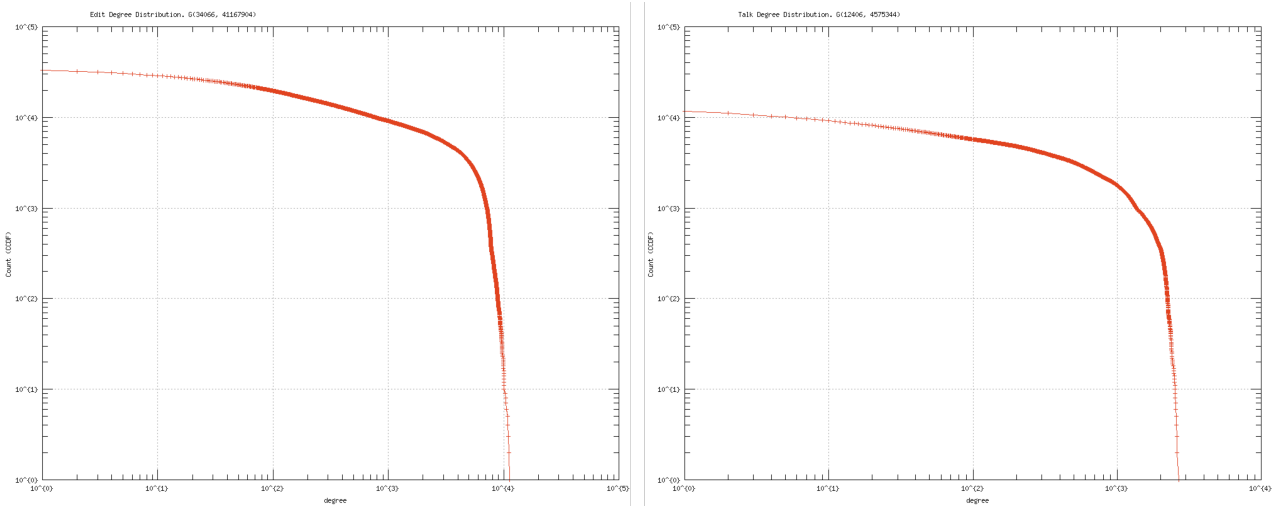


**Figure 2.** The figure shows a complementary cumulative distribution function of the edit (left) and talk (right) networks on a log-log scale. In each graph, the x-axis gives a degree k and the y-axis shows the number of nodes with degree at least k.

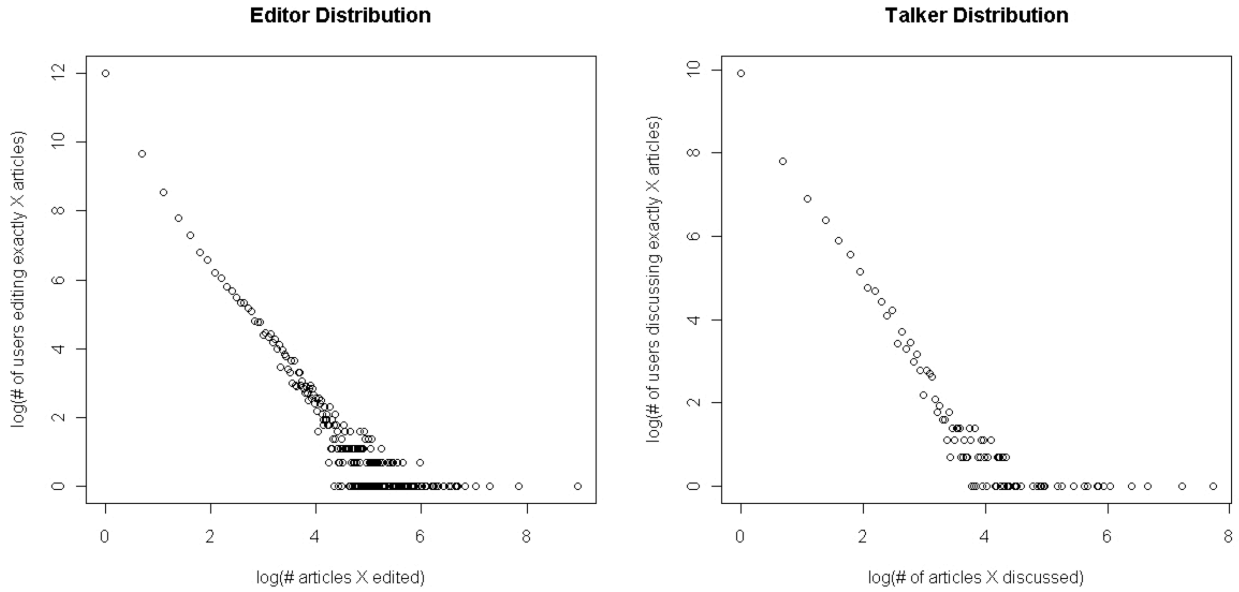


**Figure 3.** The figure shows a log-log scale distribution of editors for the Wikipedia edit and talk networks. In the Editor Distribution graph, the x-axis is the number of articles and the y-axis is the number of users that edited that number of articles. Similarly, the Talker Distribution graph shows the number of articles discussed plotted against the number of users that discussed that many articles. In each graph, we see that most users edit or discuss only a few articles, while some dedicated users edit or discuss many articles. We do not confirm whether or not the users editing and the users discussing are the same users.

In these plots, the x-axis represents a degree k and the y-axis shows the number of nodes with degree at least k. By analyzing the degree distributions in our graphs, we see that in both cases, there are a few articles that are not very well connected in the graph, and many articles that are highly interconnected.

We believe there are a few possible explanations for this pattern of degree distribution. First, the results could indicate that there are a few articles on the fringe which are edited by minor users, while there are many central articles that correspond to those edited by the major users or moderators. Another possible explanation is that some articles are more popular than others, so these articles get edited

by more users; this in turn increases the likelihood of these articles being connected to many others.

We hypothesized that most users only edit a few articles there are a few users who edit many articles. To explore this hypothesis we plotted the number of articles against the number of users who edited that number of articles (Figure 3). The plot shows that for both the edit and talk networks, a large proportion of users only makes small contributions, while there are a few power users, who contribute to hundreds or thousands of articles, thereby confirming our original hypothesis.

### 4.3 Network Paths

Graph diameter is the maximum shortest distance between any two nodes. We computed the graph diameter for both the edit and talk networks. A traditional definition of the effective diameter of a network graph is the minimum distance such that more than β (0.9 in this case) nodes in the graph are connected with at most distance d apart. However in this paper we calculate the effective diameter by a function, which is defined as the linear interpolation between the points of a hop plot. Thus, when the function is equal to β, this function gives the effective diameter [9]. A hop plot (Figure 4) is a graph showing the percent of nodes in the graph that are reachable within h hops. They give us a sense of how quickly the neighborhoods of the nodes expand with distance. The hop plots in Figure 4 show that most nodes are reachable within 3 hops which is an indication that the graph is relatively densely connected. We found the effective diameter of talk to be 3.8 and the effective diameter of edit to be 3.61. The small diameter for the Wikipedia graphs shows that most editors are active across many different articles and categories. The effective diameter also confirms the fact that most users collaborate with a lot of other users when editing or discussing an article.

### 4.4 Components

As another measure of connectivity in the graph, we calculated the percentage of nodes in the largest strongly connected component of each network. We found that the largest strongly connected component contained 94% of the nodes for the edit graph and 86% of the nodes for the talk graph. This means that for any given pair of articles in Wikipedia, we can mostly find a trail of editors from one article to the other.

### 4.5 Clustering Coefficient

We also computed the average clustering coefficient for each of the networks, which is a measure of the extent to which the nodes in the graph tend to cluster together. We found the clustering coefficient for the edit graph to be 0.82 and found the clustering coefficient for the talk graph to be 0.88. Given the high clustering

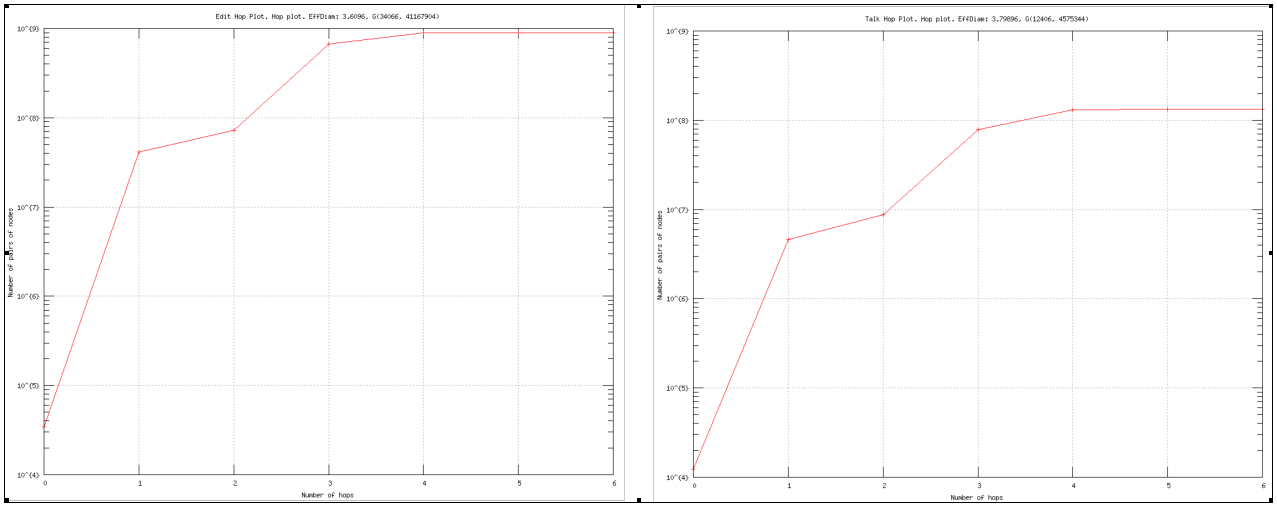


**Figure 4.** The figures show the hop plot for edit (left) and talk (right). The x axis represents the number of hops and the y axis, in log scale, represents the number of nodes reachable in h hops.

coefficients in the graphs along with the diameter data analyzed in Section 4.3, we find that both the talk and edit networks show properties of a small world network, where any one node in the graph is never more than a few hops from any other.

# 5 Communities

In this section we present our methods for identifying and evaluation community structure in the edit and talk networks as well as the results of our community detection in these networks.

### 5.1 Identifying Communities

We used the Clauset-Newman-Moore algorithm from the Stanford Network Analysis Platform (SNAP) [10] library to identify communities in our edit and talk networks. The algorithm takes as input a file in which each line gives the names or ids of two nodes that are connected by an edge and gives as output a list of the nodes in the network along with the community to which each of the nodes belongs.

### 5.2 Community Evaluation

We evaluated the communities generated by the CNM algorithm in a number of different ways. First, we examine the metric Q, which is a standard metric for evaluation of community modularity initially proposed by Girvan and Newman [5]. Q is given by the following equation:

$$Q = \sum_i (e_{ii} - a_i^2)$$

Here $e_{ii}$ refers to the fraction of all edges that lie completely within a group *i* and $a_i$ refers to the fraction of all ends of edges that are attached to vertices in group *i*. Thus, $a_i^2$ reflects the fraction of edges that would connect vertices within group *i* if the edges were to be selected at random. In this sense, Q is a measurement of whether the community division gives more within-community edges than we would expect by random chance. Values of Q between 0.3 and 0.7 typically indicate significant community structure.

Second, we analyze the resulting communities qualitatively to see if we can identify commonalities among articles that are placed in the same community by the CNM algorithm, most notably with respect to the category of these articles.

### 5.3 Results

For the talk network, running the CNM algorithm for community detection resulted in 127 communities with a Q value of 0.46, which indicates significant community structure. The average number of nodes in a community is therefore 97.69.

We further evaluated the communities qualitatively by analyzing them individually and comparing their characteristics. We found that several communities were highly cohesive with respect to categories, while a few others were not. For example, we identified one community that solely contained articles written about individual radio stations and another community with articles describing Basque people and places. Yet another community only contained articles about people, but we were unable to identify any more specific connection among these articles. By contrast, some of the less cohesive clusters contained articles with a wide range of subjects. The articles in this cluster include “The Holocaust”, “Monopoly”, and “Cisco Systems”, among many others.

The edit data showed significantly less community structure. The CNM algorithm divided the network into 88 communities with a Q modularity value of 0.21, which does not indicate significant community structure.

Additionally, through a qualitative examination of the individual communities, we were only able to identify a handful of small communities that appeared to have consistency with respect to categories. For example, we found a small community (39 articles) with articles on the topic of digital television in various countries and another small community (9 articles) on the topic of optimization software libraries. However, these communities are miniscule in light of the fact that the communities averaged approximately 387 articles in size. Thus, we were not able to find any major communities that appeared to adhere to any one particular category.

Given the very high average node degree in the edit network (1,173 edges per node), the fact that we did not find strong communities in the edit data set may not be particularly surprising. Recall that in our network each node represents an article and each edge represents the fact that at least one user has edited both articles. We decided to try an additional community detection experiment in which we imposed a stricter condition on the edges. To create this new network, we removed all edges from the edit network from where only one user had edited both articles, leaving only edges between articles where at least two users had edited both articles. We then dropped all nodes that were no longer connected to any other nodes. This representation has two useful properties: it is much sparser than the original network and it should theoretically keep the closest article relationships intact, since those are the pairs of articles where the same user edits both articles. The new network contained 18,170 articles with 1,609,584 edges between them, resulting in a new average node degree of 88.6. Running the CNM algorithm on this network yielded 537 communities with a Q value of 0.17. Surprisingly, the Q value got worse as we restricted the network to edges that indicated closer relationships. One possible explanation for this result is that our sparsification of the graph resulted in a number of node pairs that were connected by an edge, but were not connected to any other nodes. These node pairs greatly increased the total number of communities, but may not have helped to improve the modularity significantly.

To interpret the results of community detection in our data set, it is important to think about what strong community structure actually means in the context of this network. Strong community structure in general means that the network can be divided into groups within which connections between nodes are dense, but outside of which connections are much sparser. In the case of the Wikipedia networks, this would mean that discussion or editing connections within groups would be dense, but editing between groups would not be as common. Thus strong community structure might indicate that authors typically tend to edit or discuss articles that are related in some way. By also looking at the communities found in the talk network, we see that authors form communities based on subjects and often discuss the articles that are related by subject. However, in the edit network we don't see any community structure and correspondingly can note that articles in communities are generally not related by category. This seems to indicate that there are many users in Wikipedia who edit articles across a variety of different topics or categories; however, the actual discussion of articles seems to be more related along the lines of the subjects of those articles.

## 6 Conclusions and Future Work

There are a number of possible directions for future work. For this paper we were using a machine with 17.1GB of main memory, which limited the size of our edit graph to articles ranging in three categories with a total of 3230 subcategories. By growing the graph to include other categories, we think that we can get a better estimation of the existence of communities.

Another possible direction is to look at different types of relationships between articles where an edge is created between two articles when one article links to another. We can then run the community detection algorithm on this graph and see if the communities match the categories that the articles fall into or see how the link structure correlates with the edit structure that we follow in this paper.

We could also analyze the network over time and see how metrics like the effective diameter or the degree distribution change over time. This will allow us to see if the growth of Wikipedia in terms of users results in a tighter or loosely knit communities. Additionally, we might examine how the generation of content or edits around particular articles or categories of articles corresponds to the timing real world events.

We have shown in this paper that network analysis methods can provide insights into the dynamics of content creation on Wikipedia. By analyzing the distribution of users and the articles they edit, we found that many users edit

only a few articles, while a few users edit hundreds or thousands of different articles. By analyzing path lengths through the network and the clustering coefficient, we noted that the Wikipedia edit and talk networks both follow the properties of small world networks. Finally, by examining community structure in the networks, we learned that users edit across a wide variety of categories, but are more likely to discuss articles that remain within a particular category. Such insights might be used to improve or reinforce these patterns of content creation in the future.

## References


1. Wikipedia. http://www.wikipedia.org.
2. A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Physical Review E, 70:066111, 2004.
3. Leskovec, J., Adamic L.A., and Huberman, B.A. The dynamics of viral marketing. In *Proc. 7th ACM Conference on Electronic Commerce,* 2006.
4. Kwak, H., Lee C., Park H., Moon S. What is Twitter, a Social Network or a News Media? In *Proc. 19th International World Wide Web Conference,* 2010.
5. Newman, M.E.J., Girvan, M., Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004.
6. Lancichinetti A., Fortunato S., Community detection algorithms: a comparative analysis, *arXiv:* 0908.1062, 2009.
7. Newman M.E.J., Girvan M., Community Structure in Social and Biological Networks. *Proc. National Academy of Sciences*, Vol. 99, 7821-7826, 2002.
8. Hart M. Wikipedia Categories Project. http://sourceforge.net/projects/wikicategory/
9. J. Leskovec, J. Kleinberg and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.
10. Radicchi, R., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D., Defining and identifying communities in networks, *Proceedings of the National Academy of Sciences.* USA, 101:2658–2663, 2004.
11. Newman, M.E.J., Fast algorithm for detecting community structure in networks, Phys. Rev. E 69, 066133, 2004.
12. Stanford Network Analysis Platform (SNAP). http://snap.stanford.edu