A Network Modeling Approach to Assisting Collaboration in Large Scale Online Environments

Ahmed M. Mohamed
St. Cross College
University of Oxford

Supervised by: Dr. Vasile Palade

A dissertation submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science

Oxford University Department of Computer Science

September 2011

Abstract

This thesis presents a model for automated ranking and assistance of collaboration in online networks, using Wikipedia as a case study. The rise of massive collaborative networks, like Wikipedia, shows tremendous potential for harnessing collective intelligence of a crowd of people, however, its highly decentralized nature, atypical of organizational cooperation, is posing a lot of challenging questions on how collaborative networks may be structured in a manner that allows robust and effective cooperation without compromising the degree of choice of the collaborating agents.

The thesis aims to contribute to a better understanding of how collaboration dynamics in online decentralized networks impact the outcome of that collaboration. This is done by studying the collaboration network of Wikipedia articles, in terms of both the roles people play in the network and their relationships to each other.

The ranking model we developed is evaluated by comparison with existing rankings of Wikipedia articles as *Featured* and *Non-Featured* articles. The collaboration assistance model is evaluated by comparing its correlation with articles going through an improvement at some point in Wikipedia. Analysis of results indicates a few insightful properties of *Featured* and *Non-Featured* article editing networks, and of the impact of editor roles on the qualities of both outputs.

Acknowledgements

I would like to thank my supervisor, Dr. Vasile Palade, for his consistent guidance and encouragement, without which the completion of this thesis would not have been possible.

I would also like to thank Professor Bernie Hogan, from the Oxford Internet Institute, and Professor Mason Porter, from the Mathematical Institute, for the insightful discussions and remarks they both provided me with.

Most importantly, I would like to thank my family, who gave me the support and encouragement for working through that thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancement of different communication technologies, like the Internet, and the subsequent continuously pervasive adoption of these technologies by people all over the globe has allowed large groups of people to share ideas, interact and collaborate in new ways, creating global success stories, like Wikipedia, which allowed the harnessing of global knowledge towards the creation of an online encyclopedia fully managed by unpaid volunteers. The Internet has also become an incubator of preexisting networks of interaction, like academic and professional networks.

The resulting abundance of detailed human interaction network data on the web leveraged the potential for deeper quantitative analysis of complex patterns of human interaction, traditionally limited to the scope of social sciences. This potential has caused a surge in social network analysis research over the past decade, leading to tremendous insight into key properties of human networks and the development of powerful models for manipulating these networks.

On another note, the rise of massive scale collaborative networks, like Wikipedia, shows tremendous potential for harnessing collective intelligence of a crowd of people towards tackling challenging problems, however, its highly decentralized nature[22], atypical of organizational cooperation, is posing a lot of challenging questions related to how online collaborative networks may be structured in a manner that allows robust and effective cooperation without compromising the degree of choice of the collaborating agents.

1.1 Collaboration Assistance

Our thesis aims to employ a few key advancements in social network analysis and modeling to resolve the aforementioned decentralized collaboration challenge, and to further harness the tremendous power of online collaboration through proposing a machine learning framework for automated assistance of human collaborators in an online environment, in that framework; computational analysis plays the role of streamlining collaboration of human agents rather than mimicking the behavior of these agents. This will allow collaboration environments to maximize on the tremendous power of large scale distributed computational capacities for improving collaboration outcomes without needing to possess an in-depth perspective of the context of that collaboration, thus allowing it to generalize flexibly to different collaboration contexts.

Our method relies on using readily available records of Wikipedia article edits and article classifications to identify the roles of different editors, create a network relating editors of the article establish a ranking for that article based on that network and suggest potential means to remedy the quality of the article. For example, our model learns that *Featured* articles often contain more *Topic Contributors* than *Non-Featured* articles, who are focused on a very specialized topic, and that *Featured* article editing networks contain more diverse relationships than *Non-Featured* article editing networks. These properties are inferred automatically by extracting key network parameters from article editing networks, and using them as features in learning an article classification model.

1.2 Evaluation

Since the application of our model is two-fold, the first being automated rankings of articles and the second being automated assistance by recommending editors to join the editing network, our evaluation means are two-fold as well. While both are built on a set of articles in the Wikipedia categories *Social Sciences*, *Natural Sciences* and *Arts*, automated ranking is evaluated by testing the ranking accuracy relative to actual rankings of a sampled set of articles. As for the assistance component, it is evaluated by choosing a subset of articles that have gone through an improvement in ranking, and measuring the correlation of the changes in its

editing network with the suggested changes by the editor recommendation algorithm. Results of our evaluation process are encouraging, with editing recommendations showing a promising level of correlation with actual results. The advantage of combining network analysis techniques with role detection are also demonstrated experimentally.

Although Wikipedia's contribution policies are currently restrictive to direct inclusion and evaluation of the model in the network, however, this model can be used to construct an independent Wikipedia assistance tool which editors can use to decide on which articles most needs their contributions at the moment.

1.3 Key Contributions

Our model is built on recent advances in human network analysis research, that shed the light into unique features of human networks, relating to its structure, its evolution and the distinct properties it demonstrates in both, and it is also motivated by the simultaneous spur of online networking environments which are yet to realize their full potential.

Our means to detecting roles in social networks draws inspiration from emerging web based individual influence detection methods. As for our collaboration assistance model, it works along the lines of various network modeling approaches, which are built on matching a large network to a set of key exemplary parameters. For example, when building a model for evaluating the "health" of a particular editing network, it is important to include aspects like who knows whom, how close they are together and how many communities they have, these aspects translate into metrics *like connectivity, clustering coefficient* and *homophily (See Chapter 3).*

Despite drawing inspiration from all these areas, our approach is in a sense more holistic, as it attempts to address the fundamental question of how the complete connection patterns in a collaborative network relates to the output of that network. That being said, this thesis proposes three key contributions. First, automated role detection in large scale collaboration networks allows greater utility than human mediated means of role detection as presented in [1] and [2]. Second, offering a content free article ranking model provides adaptability and scalability

of ranking not afforded by ranking models using content as a fundamental block. Furthermore, this is the first attempt, to the best of our knowledge, in ranking collaboration output strictly by the network patterns of its collaborators.

Third, the proposed collaboration assistance model both presents a novel paradigm for boosting large scale collaboration in crowd sourced environments and novel means for utilizing a network modeling approach entitled *Kronecker Graphs*.

Automated Role Detection

Prior role modeling efforts, whether for Wikipedia or other networks, relied on manual intervention. This requirement poses great limitations on the ability to scale such efforts to analysis of collaborator roles in very large scale collaboration networks like Wikipedia. Through clustering editors under different natural role clusters based on statistical and structural features of these editors, we are able to automatically detect common types of general roles adopted by editors or mixtures of roles with varying degrees. Furthermore, detection of roles in that manner allows adaptability to behavior evolution, an evolution that falls extremely short of manual detection methods which are constrained to small scale or possibly obsolete role distributions in collaborative environments

Content Free Contribution Ranking

The richness of Wikipedia activity information openly available for all contributions in all languages offers deep insight into the potential quality of the content being generated. However, this dimension of Wikipedia is highly underutilized for better management of the editing process due to the scale and rate with which information is constantly being incremented into the network and the consistent evolution of behaviors and roles in the network. This requires an equally scalable means of analyzing the content and the contributions in their entirety, a facility which is many times highly evasive of manual application by human contributors. Using the article editors' networks of Wikipedia to rate content generated by it, without direct analysis of that content, needs novel methods for extracting information from these networks and relating them to the resulting quality of an article. Our method tackles that task by combining supervised

machine learning and network modeling approaches to create a robust article ranking model.

Content Free Collaboration Assistance

Given their relatively recent rise, large scale collaborative networks remain highly underutilized for harnessing collective intelligence of a large crowd of people to redefine organizational norms or solve global scale problems. That utilization could potentially be weakened by the high degree of decentralization inherent within such collaboration networks, which acts as an obstacle to ensuring accuracy and robustness of the collaboration output from these networks. Content specific means for aiding that collaboration process, like adding artificial agents for automation of parts of the collaboration process, is not always a convenient option, , due to current limitations of artificial intelligence techniques. For example, many collaboration activities require a profound level of natural language understanding, which has still not been mastered by AI techniques. By utilizing the network modeling approach of *Kronecker Graphs[3](See Chapter 3),* to relate the network structure of group collaboration to the quality of the output of that collaboration, we are effectively creating a content free collaboration assistance model which attempts to bypass the current limitations of AI in dealing with textual content.

1.4 Thesis Overview

The subsequent parts of the thesis are organized as follows; Chapter 2 justifies choosing Wikipedia as a case study, describes in a guided manner main elements of its collaboration environment and previous related work in modeling roles, ranking articles and assisting improvement of Wikipedia content. Chapter 3 explores network analysis and machine learning concepts and models employed in the thesis. Chapter 4 explains thoroughly our collaboration assistance model, with its 3 main parts, *Role Modelling*, *Article Ranking* and *Editor Recommendation*. In Chapter 5, we present implementation details, including data collection, testing baselines and evaluation metrics. In Chapter 6, we present the results of the aforementioned implementation and use these results to analyze general properties of Wikipedia collaboration networks. Finally, in Chapter 7, the main contributions of the thesis are reiterated and potential future research directions are discussed.

Chapter 2

Wikipedia as a Case Study

This chapter discusses motivations for using Wikipedia in specific and describes the Wikipedia collaboration environment in a guided manner. Next it discusses recent work in modelling roles and ranking articles in Wikipedia.

2.1 Why Wikipedia?

Since the broad theme of the dissertation is how the network structure of collaboration of a group of people relates to the resulting quality of that collaboration, the dataset used for tackling that theme must fulfill the following needs:

- Provide detailed recording of each collaborator's contributions which allows analysis of that collaborator's role in the network.
- Provide detailed recording of the evolution of the collaboration outcome over time.
- Provide adequate evaluation of the quality of the collaboration outcome.

Wikipedia and research publications come to mind when considering these data requirements, as they both specify different means for detecting collaboration of people, in Wikipedia, it could be discussion edits or article edits, and in research

publications, it could be citations and co-authorships. Moreover, there is a visible product of group collaboration in both cases, where in Wikipedia, it is encyclopedic articles, and in research publications, it is namely the publications themselves and the impact they have on the research arena. Despite the initial bias being more towards analyzing research publication networks, due to the more broad scale of impact of research publications, however, our decision was to focus on Wikipedia, mainly due to the complete edit details readily provided by Wikipedia for direct analysis. Moreover, the fact that articles are classified by quality, ranging from *Featured* Articles to *Regular* Articles (more recently including ratings for objectivity, accuracy and completeness), allows us to learn the relationship between a networked pattern of collaboration and a measurable quality of outcome.

2.2 Wikipedia collaboration platform

To provide a holistic description of the Wikipedia eco-system is an evasive task as the platform is expanding at an increasing pace and constantly evolving new forms of interaction and contributions, as mandated by the users themselves. Therefore, this section provides a guided overview of the key elements of the Wikipedia community that most relate to our proposed collaboration model.

2.2.1 Page Types

Different page types in Wikipedia are formally designated as namespaces. The following are the Wikipedia namespaces of direct relevance to our thesis and their corresponding type of members.

Main (*Designated as namespace 0)* It contains all the articles in Wikipedia. Articles are generally open for editing by all Wikipedia users, unless a page is locked for editing by certain types of users. For example, articles of many former living US presidents are semi-protected, meaning that only registered and confirmed users can edit them.

Talk (*Designated as namespace 1)* It is composed of the discussion pages of Wikipedia articles.

User (*Designated as namespace 2)* It is composed of personal pages of Wikipedia users. These pages are open for access by all users. However, only the user owning the page can edit it.

User Talk (*Designated as namespace 3)* It is composed of discussion pages of Wikipedia users and it is open for editing by all users. Editing patterns in that namespace can give indications of users who assist in the social facilitation aspects of Wikipedia, like welcoming new users and discussing projects.

Wikipedia (*Designated as namespace 4)* It is composed of current Wikipedia projects. Projects are normally started by users to improve procedural aspects of Wikipedia or to resolve problematic issues in it. For example, the aforementioned semi-protected articles policy was started as part of a project designated *Protection Policy*, in an attempt to suppress vandalism of Wikipedia articles. Like the *Main* namespace, the projects namespace is generally open for editing, unless assigned a particular protection level.

Wikipedia Talk (*Designated as namespace 5)* It is composed of discussion pages of current Wikipedia projects, and it is open for editing by all Wikipedia users. Both Wikipedia and Wikipedia Talk namespaces are indicators of users active in enhancing Wikipedia content indirectly, through organizational means.

Category (*Designated as namespace 14)* Each article is designated with one or more categories, which are not necessarily of direct relevance to each other. For instance, the article discussing Beethoven contains the categories *1770 births, German composers* and *Smallpox survivors* among others. Each of these categories is part of the category namespace. Article categories help in establishing semantic relationships between different articles in a straightforward manner, without direct analysis of the actual content of each article.

For each sampled Wikipedia user, edits falling under namespace 0 to 5 will be collected, as they capture the collaboration, user interaction and project facilitation aspects of the encyclopaedia. As for the category namespace, its categories will rather be used to relate articles together when creating a *Contributions Network* of a user.

2.2.2 Editor Types

An important facet of Wikipedia is the variation in the types of users, in terms of access levels, formal tasks adopted and identifiability. Access levels identify what kind of functions a Wikipedia user is allowed to execute, like editing protected articles and reverting pages to a previous state. Formal tasks vary from managing the voting process to administering different projects in the Wikipedia namespace like *Articles for Deletion or Articles for Improvement*. Users also vary by identity, where editors can either be IP/anonymous editors, regular editors or robot editors, commonly referred to as Bots, who take on repetitive tasks like formatting and organizing article structure.

Since our focus is on the direct impact of an article editor on the quality of the article, we will assume that access levels and formal tasks are of less importance to us than the issue of editor identity. Unlike regular users, IP's do not uniquely reflect an identity of a person, and Bots bear little relevance to the actual content present in an article, as such IP editors and Bots should be either discarded or treated in a manner different from regular editors.

2.2.3 Article Ranks

The Wikipedia community has created formal means for dividing Wikipedia articles in terms of the quality of their content. As is the case with all the previous facets of Wikipedia, these means are increasingly becoming more complex over time. However, at the top level of these classifications, articles can be broken down from best to least into *Featured, Good* and *Regular* articles. The Wikipedia community identifies the standards of *Good* articles as well written, broad in coverage, accurate, verifiable, neutral and stable, *Featured* article standards demand comprehensiveness and being well researched as additional criteria.

The Wikipedia community manages the promotion process of articles using these standards. Currently, Wikipedia English contains 3000+ *Featured* articles and 12000+ *Good* articles. Despite the existence of other means of classifying article quality, we will assume that articles not meeting the *Featured* and *Good* standards will simply be treated as *Regular* articles.

2.2.4 Information Available

All edits made in the Wikipedia namespaces are publicly available. For each edit, Wikipedia records, edit time, page name, type of edit, edit comment and article status before and after the edit. The type of the edit is identified as either a minor edit designated by the letter *m*, or an edit creating a new Page designated by *N*, otherwise an edit can be treated as a regular edit. Since the main paradigm of our approach is its content independence, we will not consider comments or the actual page content when analyzing article edits.

2.3 Project in Context

2.3.1 Approaches to Role Modeling

By role modelling, we mean a systematic breakdown process of a group of interacting elements into distinct roles adopted by each element. Over the last decade, the increasing abundance of detailed recordings of human interaction, like emails, social networks and wikis, has generated a large surge in automated role detection research. While not standardised classification of role detection approaches has been made, they can be broken down into statistical approaches and structural approaches

Statistical approaches McCallum et. al (2003) adopts a content aware approach for detecting roles through the Enron emails dataset**,** by relating the roles of people to keywords they frequently use when communicating with others. Many other general role detection approaches follow the same path. However, the most relevant one for the purpose of our thesis was conducted specifically for Wikipedia by T.Welser et.al (2011) in which he follows a semi-automated process for finding social roles in the Wikipedia collaboration platform. He implements a top down approach for role construction, by defining four specific roles to look for, namely, vandalism fighters, technical editors, experts and networkers [1]. Then, users fulfilling these roles are sampled manually and their edit statistics in namespaces 0 to 5 (See 2.2) are used as signatures to assign a group of sampled editors to these roles. While the assignment is purely statistical, Welser also qualitatively analyzes structural properties of each user based on his editing pattern in user talk pages.

Structural approaches On the other hand, relatively little research has went into using structural properties of a user to assign a role to that user. An exception to that is Shi et. al. (2010) who followed a structural approach to analyze citation networks, constructing six different structural metrics to distinguish different kinds of researchers. They posited that researchers can be divided into idiosyncratic citers, whose citations exhibit a weak relationship to each other, within community citers, whose citations normally come from a highly focussed topic, and brokerage citers, whose citations are weakly related by a large common topic[4]

2.3.2 Approaches to Article Ranking

The tremendous pace with which the Wikipedia article base has been growing, adding to that the temporally unstable nature of Wikipedia articles due to its open collaboration platform, has led to an increased interest in research of automated ranking methods for Wikipedia articles. Some research efforts build on general document ranking paradigms and others employ the rich editing structure of Wikipedia for ranking its articles. Here we focus on reviewing the Wikipedia specific approaches.

Wikipedia Specific methods for automated ranking of Wikipedia articles do not fit one stereotype; however, they all somehow build on the rich collaboration information of Wikipedia articles. The earliest and simplest approach was put forth by Lih (2004) who derived a correlation between article rank, and number of edits and editors [5]. Stvilia et. al (2005) addressed the ranking issue by mixing a group of editor role sensitive metrics like editor authority, age and administrator ratio with metrics like content diversity and readability [6]. Adler (2007) ranked articles by tracking each word to its contributors [7]. He scored each word by the trustworthiness of its contributor, which is a function of his edit survival time. Then, he combined all the words to give a respective score to the article as a whole. A more distinct approach was proposed by Wöhner et. al (2009) , who analyzed the editing life cycle of articles and broke down the article's content into persistent and transient contributions, then used that break down to rank the article [8]. Notably, Blumenstock (2008) proposed a large set of simple metrics like word count, sentence count and external links giving 97% classification accuracy in distinguishing between *Featured* and *Non-Featured* articles[9]

The high classification accuracy of Blumenstock's simple approach might imply lack of a need for a new more complex approach for classifying Wikipedia articles. To put Blumenstock's results in perspective, Wikipedia English currently contains less than 4000 articles constituting almost 0.1% of the total article count. As for the remaining articles, the average size of articles is only 500+ words [19]. As such, a 97% accuracy level on a balanced sample of Wikipedia articles is both unsurprising, and of limited use on its own, given the very low percentage of Featured articles in random Wikipedia article samples. Moreover, it fails to distinguish Featured and Non-Featured articles of comparable size. As such more nuanced classification methods are needed to rank articles of comparable sizes, and that is precisely the approach we chose to adopt.

Chapter 3

Overview

This chapter provides an overview of key network analysis and machine learning concepts and terminology employed throughout the thesis.

3.1. Social Network Analysis

This section first presents an overview of the key network terminology and network concepts employed in the thesis. Then, it goes on to discuss key properties and network modelling approaches for real networks.

3.1.1 General Network Concepts

A real world network is commonly represented as a graph. A graph $G = \{V, E\}$ is defined in terms of vertices V and edges E, where each node represents a distinct element in the network and each edge $E_{ij} = \{v_i, v_j\}$ distinctly connects two vertices v_i and v_j. Vertices and edges can also be referred to as nodes and connections respectively.

Weighted Graph: A weighted graph is one where an edge $E_{ij} = \{v_i, v_j, w_{ij}\}$ is additionally represented by weight w_{ij} between vertices v_i and v_j.

Adjacency matrix: an adjacency matrix A is a $|V|\times|V|$ matrix representing a graph G, where $|V|$ signifies the number of vertices in the graph, where $A_{ij}=\{1\}$, if there exists an edge E_{ij}, and 0 otherwise. As for a weighted graph $A_{ij}=\{w_{ij}\}$ for an edge.

Node degree: a degree d of a node v_i signifies the number of nodes v_i is connected to.

Self-edge: an edge connecting a node n_i to itself i.e. $A_{ii}=\{1\}$

Cross-edge: an edge connecting two different nodes.

Triad: three nodes v_i, v_j and v_k form a triad if at least two edges exist between these nodes. If three edges exist between these nodes, we call it a closed triad**.**

Directed network: a directed network is one where an edge is both defined by the pair it connects and the direction of that connection, i.e. in a directed network, an edge $E_{ij}=\{v_i,v_j\}$ explicitly signifies a connection from vertex v_i to vertex v_j.

Undirected network: similarly an undirected network is one where an edge $E_{ij}=\{v_i,v_j\}$ signifies a connection from any of the two vertices to the other.

Bipartite network: a bipartite network is one which can be divided into two groups of nodes, such that edges only exist between nodes of different groups. Typically, bipartite networks relate nodes of two different classes. As related to Wikipedia, a group of articles and their categories can be seen to represent a bipartite network.

Scale free network: a scale free network is one which exhibits self-similarity of its macro and micro components [13]

Editing Network: Throughout the thesis we will use that expression in reference to the network constructed from editors of an article by connecting two editors if they satisfy a certain arbitrary relation.

Contributions Network: we will use that expression to refer to the networks constructed from articles edited by one editor by connecting two articles if they satisfy an arbitrary relation.

Homophily: the tendency of a node to connect to nodes with similar network characteristics to itself.

Heterophily: the tendency of a node to connect to nodes with different characteristics from itself.

Node Community: A group of nodes with relatively high intra-connectivity and relatively low connectivity to other nodes outside of the group

3.1.2 Network Metrics

To be able to accurately analyze real world networks and distinguish them from randomly generated networks, many network analysis metrics have emerged. Below is a review of some of the metrics of potential usefulness for our thesis.

- **Node Betweeness Centrality*:*** Betweeness of a node V_i is the fraction of all shortest paths between any two other nodes in the network which go through that node **,** as given by the following equation;

$$Bet(v_i) = \sum_{j,k} \frac{SP_{jk}(v_i)}{SP_{jk}}$$

;where $SP_{jk}(v_i)$ are the number of shortest paths between any 2 nodes j and k passing through node v_i and SP_{jk} are the total number of shortest paths between j and k.

- **Node Closeness Centrality*:*** It is the mean distance between a node and all other nodes in the network.

  Practical use: The two previous metrics reflect in some manner the importance of a node in a network.

- **Clustering Coefficient:** The average percentage of closed triads between connected triples in the network. It is computed by the following equation;

$$C = \frac{1}{|V|} \sum_{v_i} C(v_i)$$

; where V are the nodes in the network and for every node v_i, $C(v_i)$ is given by the following;

$$C(v_i) = \frac{Closed\ Triads\ Connected\ To\ \mathrm{v}_i}{Triads\ Centered\ on\ \mathrm{v}_i}$$

Practical use: Larger clustering coefficients reflect networks with high cohesiveness at the scale of small communities, i.e. a friend-of-a-friend is very likely to be my friend.

- **Connectivity:** Percentage of nodes in the largest connected component of the network**.**

- **Density*:*** The ratio of the number of edges in a network to the maximum number of possible edges in that network, as given by the following equation;

$$D = \frac{2|E|}{|V|(|V|-1)}$$

- **Diameter:** This is simply the number of hops between the two farthest apart nodes in the network.

Practical use: Higher Connectivity, larger densities and smaller diameters reflect a more cohesive network at the macroscale.

3.1.3 Properties of Social Networks

The concept of a social network is a broad concept including friendship networks, business networks, collaboration networks, citation networks and information exchange networks, among many others. Each of these networks may naturally exhibit its own distinct set of static and dynamic properties that can be captured by metrics such as those described in the previous section.

Despite that uniqueness, research conducted over the past decade, spanning from studying online social networks to phone call networks etc. has consistently

demonstrated that real world networks exhibit some key properties. Two very important properties of those relate to node degree distributions and network diameters.

Node degree distributions

The node degree distribution of a network is simply a function relating degrees of nodes to the number of nodes having such degree. Many real world networks have been found to exhibit a power law degree distribution [15], informally this means that most nodes have low degrees and as the degree increases less nodes have such degrees. More formally, a degree distribution follows a power law if the number of nodes N_d of a degree d follows this equation $N_d = Kd^{-n}$, for some constant k and exponent n [11].

To confirm Wikipedia exhibits such property we tested a full Wikipedia talk network until January 2008 comprising 2 million users, provided by Stanford Network Analysis Platform [12], where a user is connected to another if the first user edited the other's talk page. The result as displayed in Fig.3.1, confirms a power law distribution. Tests conducted on article editor networks further confirm that pattern. Power law distributions exhibit the aforementioned notion of a scale free network, which is essential when choosing the best algorithm to extract parameters of Wikipedia networks.

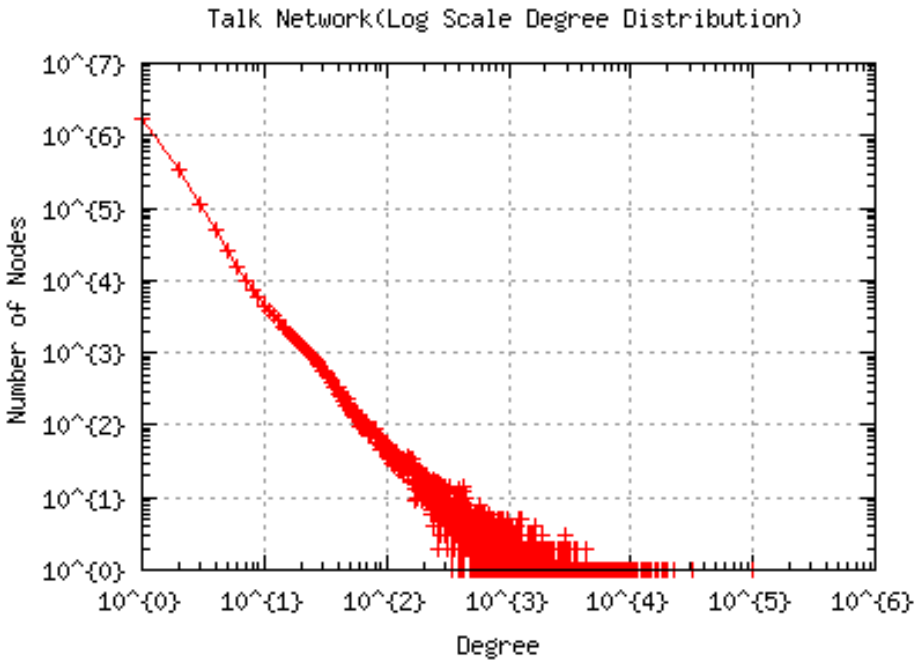


Figure 3.1 Degree Distribution of Full Wikipedia Talk network

Network diameters

Another interesting structure relating to real-world networks is what can be coined as a shrinking diameter, i.e. as the network increases in size, the diameter of the network actually decreases. This is a key feature of real world networks distinguishing them from randomly generated networks which slowly increase in diameter as they grow.

Since a network's diameter is relatively contingent upon a small number of poorly connected nodes, outlier nodes can often times give a misleading measure of network diameter. An alternative measure resolving the aforementioned issue is effective diameter (Tauro et. al., 2001), defined as the maximum number of hops needed to reach 90% of the nodes in the network [14]. That distinction is particularly useful for our case as an article's Editor Network is relatively small in size, as such effective diameter is the more useful property to look for.

3.1.4 Network Models

Over the last decade significant advancements have been made in relation to creating models that can describe and generate realistic networks. This has largely been motivated by the increasing ubiquity of large scale human networks which are available for direct analysis online. A central tenet defining the quality of any model in capturing or simulating realistic networks is its ability to adequately reflect several properties of these networks like the power law and the shrinking diameter properties discussed in the previous section.

Notable methods for modelling real world networks include the Forest Fire model proposed by Leskovec et al.(,2007), which generates connections to new nodes in a probabilistic manner similar to how a fire reaches a tree from another tree[16], and the Small-World model, which starts off by a generating a highly clustered network, then randomly connects pairs of nodes from different clusters[17]. Despite capturing many properties of real world networks, these models have not been successful in mimicking overall properties of human networks, as each model focuses on a particular property of human networks and tends to fail on others[3]

Furthermore, our aim is to a utilize a network model in our thesis for the purpose of extracting key parameters from editing networks that best reflect the unique properties of such networks. Most of these models fall short of that target as they rather focus on the generative aspect of network modelling by attempting to create networks that satisfy a set of predefined network properties. One recently introduced model by Leskovec et. al (2009) stands out due to both its ability to reverse the generative process by extracting basic parameters from existing networks and to adequately capture different network properties holistically[3]

Kronecker Graphs

The kronecker graph model follows a recursive paradigm employing what is known as the *Kronecker Product* of matrices. Formally, they are based on the following 2 definitions, as laid out in Leskovec et.al. (2005)

Definition 1(Kronecker Product)*:* *Given a matrix X of size $a \times b$ and a matrix Y of size $a' \times b'$, Kronecker Product Z of these 2 matrices is given by the following equation;*

$$Z = X \otimes Y = \begin{pmatrix} x_{1,1}Y & \cdots & x_{1,b}Y \\ \vdots & \ddots & \vdots \\ x_{a,1}Y & \cdots & x_{a,b}Y \end{pmatrix}$$

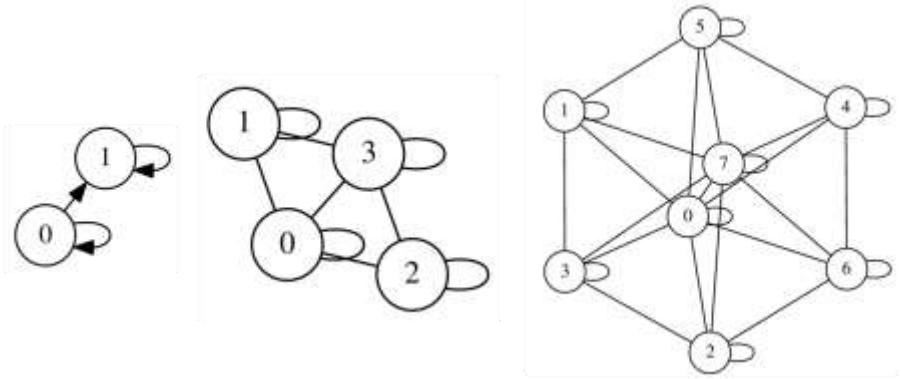
i.e. a kronecker product creates a matrix where Y replaces each location in X multiplied by this location's scalar value.

Definition 2(Kronecker Power): X_X *, The X-th kronecker power of a matrix* X_1 *is defined by the following equation:*

$$X_X = \underbrace{X_1 \otimes X_1 \otimes X_1 \otimes X_1 \otimes .... X_1}_{\text{X } times} = X_{X-1} \otimes X_1$$

By considering X_1 an adjacency matrix of a graph then X_X can be seen as the graph generated by applying the Kronecker Product $X-1$ times on adjacency matrix X_1,[23] where each Kronecker multiplication increases the size of the matrix by an exponential amount

Figure 3.2, better illustrates how Kronecker Products generate graphs and how these graphs exhibit self-similarity, in the sense that the values in the initiator matrix can be seen to represent general properties of the graph which will be prevalent in larger Kronecker products of that graph due to the recursive creation pattern. Note that node labels in the figure do not reflect actual node correspondence between different levels of the *Kronecker Multiplication* process, as each node at every multiplication is expanded into an initiator matrix. A *Stochastic Kronecker Graph* is one where the initiator matrix has continuous values between 0 and 1 and an edge in that case is present with a probability equal to the value it has after *Kronecker Multiplication*. In that case the graph in figure 3.2 would miss some of its edges if the weights were less than 1.



Graph X_1 *Graph* X_2 *Graph* X_3

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} X_1 & X_1 \\ 0 & X_1 \end{pmatrix} \qquad \begin{pmatrix} X_1 & X_1 & X_1 & X_1 \\ 0 & X_1 & 0 & X_1 \\ 0 & 0 & X_1 & X_1 \\ 0 & 0 & 0 & X_1 \end{pmatrix}$$

Adjacency Matrix of X_1 | Adjacency Matrix of $X_2 = X_1 \Theta X_1$ | Adjacency Matrix of $X_3 = X_2 \Theta X_1$

Figure 3.2. Kronecker Multiplication Example

Kronecker Graphs possess several important properties that make them suitable for our analysis purposes:

- They follow power law like degree distributions of real networks[3]
- They exhibit the densification and shrinking properties found in real networks but which remained a serious shortcoming of other techniques[3]
- Being flexibly defined by initiator matrices with variable complexity give it adaptability in dealing with real graphs of different types, for example by increasing the values of the diagonals, networks with high *homophily* can be created and vice versa.
- In addition to being used to generate realistic networks it can also be used in the reverse sense to extract parameters (initiator matrices), which best express characteristics of an existing graph. That feature of Kronecker Products will later be explored in Chapter 4.

3.2. Machine Learning

3.2.1 Machine Learning Terminology

Supervised learning refers to a class of problems where a set of data points is provided along with a label for each point and the purpose is to determine a mapping function from location to label such that new points not part of the dataset can be accurately labelled.

Unsupervised learning unlike its counterpart, in unsupervised learning, no output is provided, and the question in that case is rather how to reveal hidden underlying structure relating the different data points together by strictly relying on the features

Feature Space it is an N-dimensional vector space created by assigning each feature of the data points a distinct axis.

3.2.2 K-Means

K-Means is used to solve the common unsupervised learning problem of data clustering, i.e. how to break up the data provided into groups such that elements

within a group manifest strong similarity and elements from different groups manifest limited similarity. The center of each of the produced clusters is commonly referred to as a Centroid C, which is represented by an F dimensional vector $\vec{C}$ signifying C's location in the feature vector space.

Following specification of K, the number of clusters desired, K-Means repeats a two-stage procedure, composed of the assignment and the refinement stages. The assignment stage involves assigning data points to the cluster centroid closest to it in the F dimensional vector space, where F is the number of features used, by any measure of distance chosen. The refinement stage involves recalculation of the cluster centroids following the change of each cluster's members. The above procedure is repeated until no further change in centroid locations is possible.

Chapter 4

Methodology

Our complete algorithm for assisting collaboration in Wikipedia will be described in detail in this chapter. This algorithm is composed of three components, Role Modeling, Article Ranking and Editor Recommendation, as shown in Figure 4.1.

The role detection process is composed of two sub-components, feature construction and role clustering. In the feature construction stage, statistical and structural features are computed for each editor in a sampled group of editors. In the role clustering stage, the list of editors and features is used to construct natural role clusters under which Editors fall.

Once this clustering model is developed, all editors in an article under consideration will be assigned their respective roles. This is done for each editor independently, irrespective of the roles of other editors. A relationship network is constructed between these editors based on mutual collaborations they worked on. Then, network parameters are extracted for each article editing network using the KronFit (Kronecker Initiator extraction algorithm). These parameters along with the classification of the article are used to learn representative network parameters for each class. An input article is ranked using a classification algorithm which compares the article's network parameters to those of each class.

Following classification of articles, articles ranked as *Non-Featured* go through the Editor recommendation stage, which first calculates the needed roles in the articles, retrieves editors playing these roles and samples from them the editors which best optimise the likelihood of the article becoming a featured article.

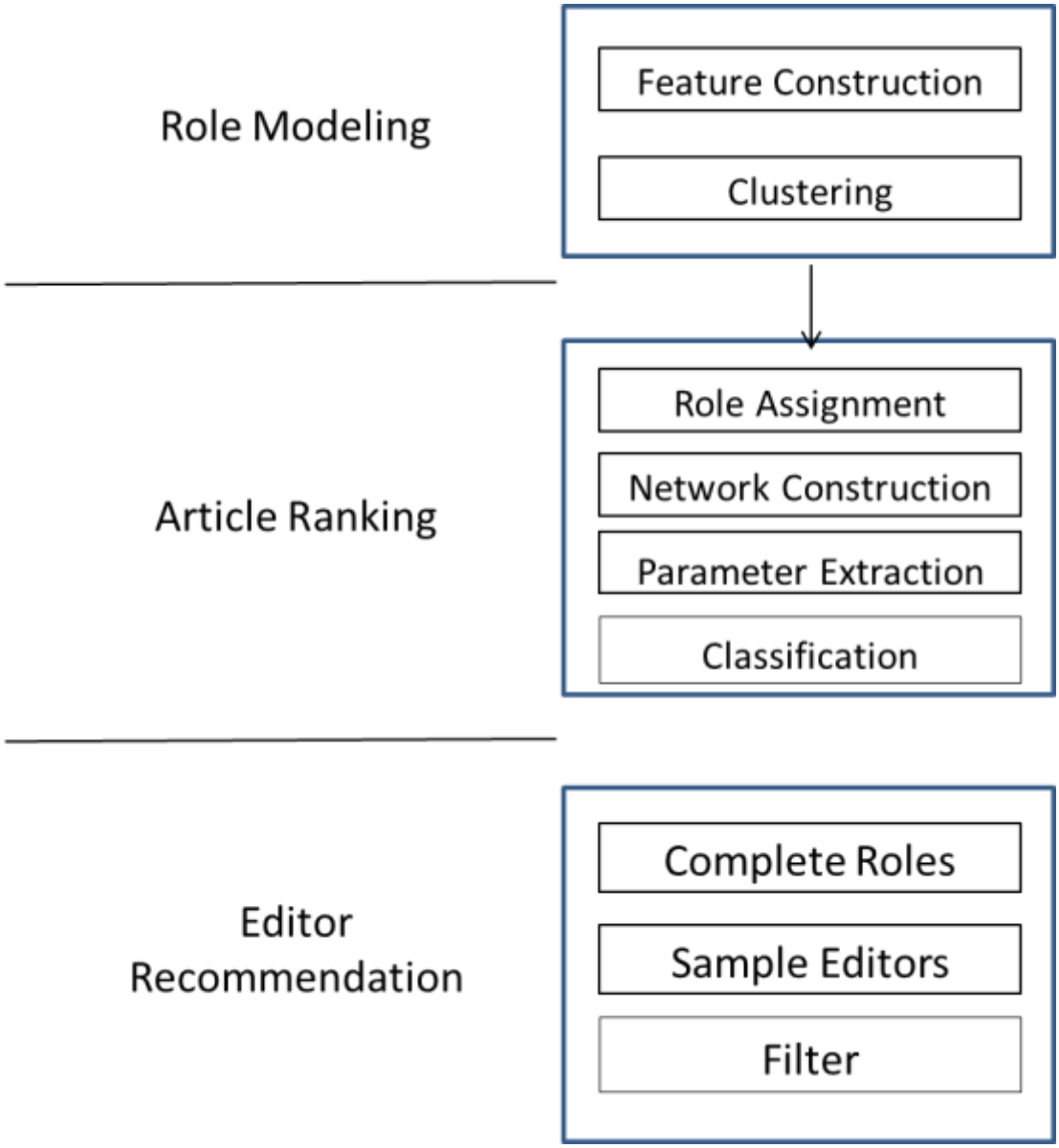


Figure 4.1. Complete Algorithm Breakdown

4.1 Role Modelling

Since editors are the building blocks of an article editing network, each with a different contribution pattern, thereby a different impact on the article, the basis on which analysis of an editing network needs to be built is the role of each editor in the network. As mentioned in Chapter 2, roles adopted by users of Wikipedia vary consistently and are increasingly diverse and often times hard to discern form quantitative analysis. For that reason, our conceptualization of editor roles will be built fundamentally on distinguishing abstract definitions of roles like proof readers, vandalism fighters, social facilitators, topic contributors and distributed contributors.

To generate a model of roles in Wikipedia, we will use a representative sample of Wikipedia editors, assuming that the general types of roles found in a sufficiently large editor sample reflect naturally recurring roles in the network. Although this assumption may not always be true, our high level perspective of editor roles minimizes possible discrepancies. Features are computed for each of these editors, which are subsequently used for grouping editors into role clusters. These clusters are the ones to be later used for assigning a role to any new editor.

4.1.1 Feature Construction

Since our main purpose is to detect editor roles in broad terms as defined before, the kind of features computed for distinguishing roles are established from a top down approach, in the sense that the measured features translate directly to certain traits indicative of the aforementioned understanding of roles. Such construction creates a bias for obtaining meaningful role distributions dictated by the human understanding of what a role entails, while not enforcing an inflexible predefinition of the roles that will emerge.

To be able to construct an adequate representation of a user's fingerprint that can lead to effectively determining his functional role in the Wikipedia article editing network, two distinct kinds of features are constructed in parallel, statistical features and structural features. The following sections will describe each of these two kinds of features in more detail.

Statistical Features

- **(M0) Article Count:** the number of distinct articles edited by a user.

- **(M1) Article Creation Rate:** the ratio of the articles created to the articles edited by a user. This distinguishes users taking up the role of creating new articles to those improving existing ones.

- **(M2) Article Creation Size:** This the average size of articles created by the user. This metric distinguishes people who expand the Wikipedia article network by creating new stubs from those who also enrich that network by creating bigger articles.

- **(M3) Article Edit Distribution**: This is the ratio of the number of article edits made by the user to the total number of articles edited by the user. Users who are focussed on particular topics would intuitively have lower distributions than those editing articles in a more general fashion.

- **(M4) Average Edit Time** $Av = \frac{T}{N}$ where T is the sum of times a user started editing a page relative to the time it was created and N is the total edits the user made. Intuitively, we would expect users who like to create and enrich new articles to have low average edit times.

- (**M5**) **Category Edit Distribution**: which is the ratio of the number of categories edited by the user to the total number of edits made by the user. This metric is expected to give similar results to M4. However, it increases robustness as users who edit a particular topic may not be editing particular articles in that topic but maybe rather dispersed in focus.

- **(M6) Contribution Size:** This is the ratio of characters contributed by the user to the edits made by the user. Proof readers would typically have a low contribution size whereas content contributors would typically have a high contribution size.

- **(M7) Correction Ratio** This is the ratio of the number of reverts made to a user's edits to the total number of edits the user makes. This should typically distinguish users who contribute low quality or opinion driven content from those who contribute higher quality less biased content.

- **(M8) Edit Frequency:** This is the rate of edits made by the users per day. The value of this metric is not particularly biased to any role; however, it's an important feature for the clustering process which may reveal a particular pattern for each cluster.

- **(M9) Main-to-Discussion:** This the ratio of the number of article edits made to the total number of edits made in either articles or article discussion pages.

- **(M10) Minor Edit Percentage:** This is the percentage of minor edits made by the user. Proof readers would typically have high minor edit percentages , whereas content contributors would typically have lower percentages.

- **(M11) Project Ratio**: This is the ratio of the edits made in the wiki namespace to the total edits made by the user. This distinguishes users who enhance Wikipedia by facilitating cooperation and making suggestions from users who focus on editing articles

- **(M12) Talk Ratio:** This is the ratio of the edits made by the user in talk pages of the main, user and wiki namespaces to the total number of edits made by the user. This metric aims to distinguish social facilitators from content contributors.

In summary, the former 11 metrics were designed to detect the following aspects of users' editing behaviour, (1) Content contribution **(M0,M1,M2, M3, M4,M5, M6),** (2) Proofreading, Vandalism Fighting and other general content improvement means **(M7, M8, M10**), (3) Editing Facilitation and Organizational duties **(M9,M11,M12)** .

Structural Features

An editor's structural features are extracted from an editor's *Contributions Network*. That network is constructed by first creating a bipartite network of articles and categories under which these articles fall, categories with more than 300 articles are removed as they present weaker relations between articles. For each of the remaining categories two articles are linked together in an article network if they had one or more categories in common. Subsequently, the following features are computed for each of these networks, these features are adaptations from Shi et. al. (2010), who follows a similar method to analyze citation networks[4]. Structural features are mainly used to distinguish between the three different kinds of editors shown in figure 4.2 namely *Topic Contributors*, focussing on a specialized topic, *Distributed Contributors*, editing interrelated topics and *Spontaneous Contributors* randomly surfing and editing articles.

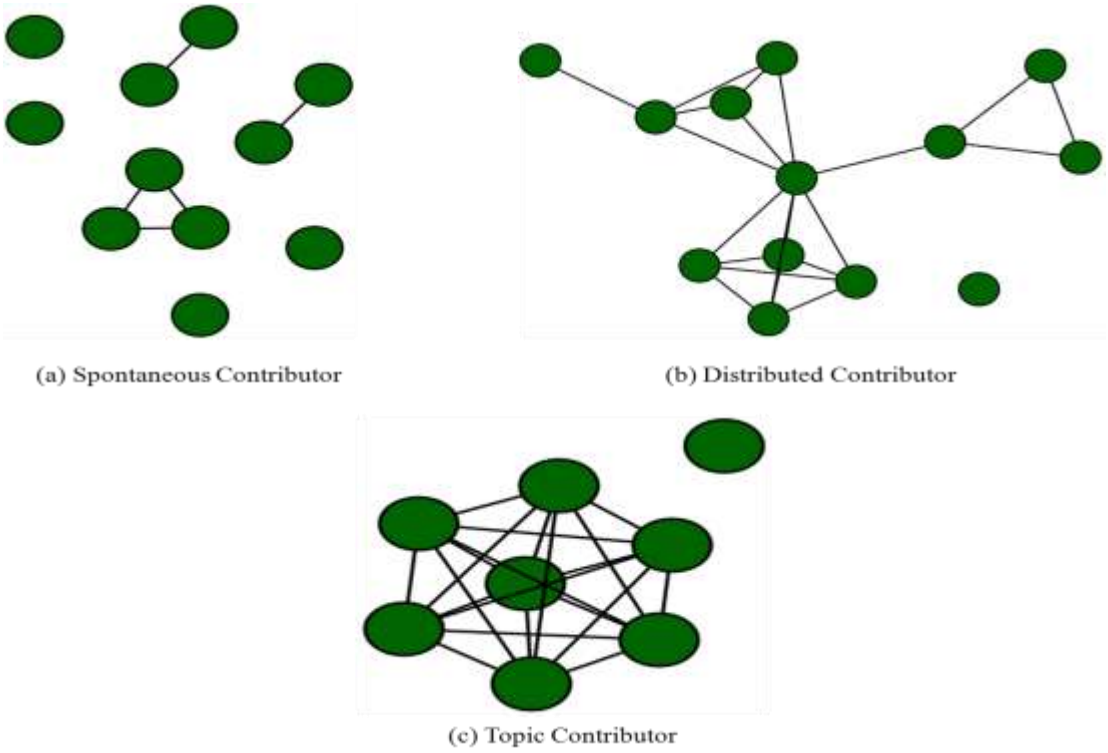


Figure 4.2. Three Contribution Networks representing distinct types of Contributors generated from collected Wikipedia dataset

- **(M13) Clustering Coefficient:** As mentioned in Chapter 3, clustering coefficient reflects the level of cohesiveness in the network at the micro scale. *Topic Contributors* and *Distributed Contributors* will have high clustering coefficient whereas *Spontaneous Contributors* will have lower ones.

- **(M14) Connectivity:** As mentioned in Chapter 3, connectivity reflects, higher network cohesiveness at the macroscale. *Topic Contributors* and *Distributed Contributors* will generally have higher connectivity values than *Spontaneous Contributors*, as their article networks have many interrelations.

- **(M15) Density:** *Topic Contributors* are expected to have high density networks whereas *Distributed* and *Spontaneous Contributors* are expected to have lower density networks.

- **(M16) Maximum Betweeness Cenrality*:*** The maximum betweeness centrality of nodes in the *Contributions Network*. As mentioned in 3.1, centrality measures reflect the importance of a node in a network. As evident from figure 4.2(b), *Distributed Contributors* will have many articles connecting different communities of articles, as such these connectors will have high betweeness, as they act as a kind of bottle neck in the network, whereas *Topic Contributors* or *Spontaneous Contributors* will have lower betweeness values.

In summary, *Spontaneous Contributors* can be detected by generally low values of all four structural features, *Topic Contributors* can be detected by generally high values of all four structural features and *Distributed Contributors* can be detected by a density lower than a *Topic Contributor* **(M15),** and a simultaneously higher maximum betweeness **(M16).**

4.1.2 Clustering

After calculating the aforementioned statistical and structural features, we use these features to perform clustering of the sampled users to extract a basic

assignment of natural roles such that these roles can later be used as a benchmark for classifying other users when analyzing an article editing process. For that reason, accuracy of the cluster centroids were the central issue driving the algorithm's design rather than accuracy of the actual clustering of a sampled dataset that would be of little relevance for our purposes later.

Motivated by that notion, we developed the K-Sampling clustering algorithm, which attempts to find natural centroids signifying distinct roles present in the Wikipedia Article Editing process. K- Sampling is an optimization algorithm that is built on the K- Means Clustering algorithm; however, it aims to detect cluster centroids that reflect natural types of clustering (natural roles in the network) while minimizing the likelihood of over-fitting the clustering centroids to the sample data set used for calculation.

K- Sampling can be considered as a Meta Learning algorithm, i.e. it embeds another learning algorithm within it as a basic component. As such, *K-Sampling* uses *K-Means* in its base. *K-Sampling* starts off by generating N samples from a dataset, then uses K-Means for creating a cluster model for each set, and then iteratively compares the cluster assignments of the intersecting points of each two models in the set, then selects the top performing M models based on their average match with all the other models. These models are then fed back as part of a new group of N samples and the algorithm is reapplied until a model is obtained which performs above a certain threshold or until a predefined maximum number of iterations is reached. A pseudocode representation of the algorithm is provided in the next page.

The rationale behind this process is to avoid creating one clustering model with potentially arbitrary centroids while not relying instead on a smaller sample to perform the clustering in its isolation.

Algorithm 1. K-Sampling Algorithm

Input: Data points D, Number of clusters K, Number of samples N, Size of a sample S, Set of Cluster Models CM, Number of Fedback models M, Set of Fedback Models FM, Maximum iteration Mx

```
FM={}
-> Start
CM <- FM
For i=1 to N
        J       <-     Sample (D,S)
        C       <-     KMeans (J,K)
        CM   <-     CM. Append (C)
End

For j=1 to Length (CM)
        For k=1 to Length(CM)
                Score(Cj)    <-      Score(Cj)    +       Match(Cj,Ck)
        End
End

For i=1 to N
        If  Score(Ci) in max M Scores
                Add Ci to FM
End

iterations <- iterations +1

If iterations >Mx
        Return  max Ci in FM
else
        Go to Start
```

4.2 Article Ranking

Following the extraction of basic roles in Wikipedia, we need to map a role to each editor in the editing network of an article. This first step to retrieve feature information about each editor in the same manner carried out for the sample editors in 3.1. Then each editor is assigned a role.

Alongside that process, we construct the *Editor Network* for each article, where each two users are connected in that network if they both edited at least another article in common. Then, basic network parameters representing an article are derived using a network modelling algorithm called *KronFit*, employing Kronecker graphs described in Chapter 2. Finally, a classification algorithm is used to determine rankings of articles, as either featured, good or regular, based on both its KronFit parameters and its role distribution.

4.2.1 Role Assignment

A straightforward approach is adopted for carrying out the role assignment process, where features described in 4.1 are calculated for an editor and based on these features; the editor is assigned to the nearest role centroid, from those initially derived in 3.1. as given by the following equation;

$$r_e = \arg\min_c \| \vec{r}_c - \vec{e} \|$$

; Where $\vec{e}$ and $\vec{r}_c$ are the vectors representing editor e's and role centroid c's location in the feature vector space.

Although, each user can naturally alternate between different roles, however, since our purpose is to analyse the general role distribution in the article, such effect can be largely discounted when taking into account many editors. Subsequently, the role distribution of the whole article is derived by a sum of all the article editors.

$$R = \frac{1}{n}(E_1 + E_2 + ... + E_n)$$

;where E_i is a one dimensional vector containing a positive value for the column signifying E_i's role *and* n is the total number of editors in the article.

4.2.2 Network Construction

As aforementioned, the editor network of an article is created by linking editors with common articles edited in the past. To avoid obtaining a dense graph, with users having a large number of edits being connected to most other users, we only consider the top 20 edited articles for each user. By doing so, we allow an edge between two users to semantically reflect the interests of these two users.

4.2.3 Parameter Extraction

To understand the nature of the network parameters extracted from the *Editing Network* of an article we need to describe the *KronFit* algorithm employed for such task. This algorithm was presented by Leskovec et al (2010) in the Journal of Machine Learning [3].

KronFit

KronFit (short for Kronecker Fit) employs the Stochastic Kronecker Graphs paradigm for modelling real world networks described in Chapter 2. However, opposite to the basic application of Kronecker Graphs which is generating synthetic networks similar to real ones, KronFit attempts to extract an initiator matrix from an existing real world network.

More formally, given a network A with N nodes, and an arbitrary initiator matrix $\Theta = N_1 \times N_1$,the task is to find the optimal $N_1 \times N_1$ values of the matrix which are most likely to give rise to network A if applied as a Kronecker Product for an arbitrary number of times, as illustrated by figure 3.2. This amounts to a natural likelihood maximization problem as given by the following equation

$$\arg\max_{\Theta} P(A|\Theta)$$

KronFit approximates the initiator matrix K_0 using both A and Θ to generate N_P network permutations of network A containing different edge connections from it.

These permutations are compared to network A and the difference between the network and its permutations is used to optimize the value of the initiator matrix Θ using gradient descent where iterations are stopped when modifications in Θ no longer improve the comparison result of the N_p permutations to A. For a more rigorous analysis of the algorithm's approach, please reference the original algorithm description in [3].

Key Values:

- N_P : Number of permutation comparisons per iteration
- $|\Theta|$: Initiator matrix size

Aside from being a linear algorithm, therefore feasible for analysis of typically large real world networks, KronFit provides an intuitive decomposition of complex networks into related parameters presenting a basic much smaller network which has been shown to reflect many of the structural features of the original networks, even at the scale of a two by two initiator matrix [3]. As such, it is useful for relating different networks together based on the similarity of their initiator parameters which are much easier to analyze than comparison of the two sizeable networks in their original states.

Applying KronFit to Wikipedia

Once we have created an editing network of an article as described in 4.2.2, we will employ the *KronFit* algorithm described above to extract the initiator parameters of that network. Note that the size of the initiator matrix can be arbitrarily varied, however, since initiator matrices as small as that with two nodes was demonstrated to describe fundamental properties of large real world networks, we will use an initiator matrix of size 3x3 and a permutation number of 1000 for efficiency limitations.

4.2.4 Classification

The aforementioned parameter extraction process in 4.2.3 and the role distribution calculation in 4.2.1 will be repeated for a sample of *Featured, Good* and *Regular* articles, generating Θ *and R* parameters of each article. Note that articles strongly

vary in number of edits and editors. For example, the number of editors of a Featured article can vary from as little as 20 to as much as 4000. To be able to properly contrast editor networks of different classes of articles, the articles sample will contain articles of comparable edit sizes across the three classes (See Section 5.1).

Following that, the Θ *and* R parameters will be combined as features representing articles. *Good* and *Regular* Articles will be merged and labelled as *Non-Featured* articles. This modification of article classes to only two allows us to perform a binary classification procedure using a *Support Vector Machine* (SVM) soft margin classifier. An SVM classifier is used, because its classification is built on the notion of increasing the distance between the two classes in the feature vector space, which fits with the fact that our classes reflect two different levels of article quality, as such the more we move towards Featured articles farther away from the classification boundary, this could be interpreted as moving more towards higher quality articles within the *Featured* articles class. Since our purpose is to properly label article quality to allow improvement of an increased portion of articles, we will optimize the SVM classifier to precision over recall of *Non-Featured* articles.

4.3 Editor Recommendation

Following classification of Articles as *Featured or Non-Featured* articles, the editor recommendation stage aims to improve articles classified as less than *Featured* quality. By starting out with the role distribution R of an article, we compute the types and number of roles needed to augment the editor network such that its modified role distribution R_A approaches the representative role distribution R_f of featured articles.

Following determination of the needed roles, we create an intermediate list R_{in}, which contains for each role, twice the number of editors needed to fill that role. Then, by executing an optimization run using the representative initiator matrix parameters Θ_f of featured articles, half the editors in R_{in} are selected for editing the article.

Note that given the extensive nature of our thesis which spans three distinct network analysis modules, our approach to recommending editors by optimizing the Kronecker initiator will follow a highly simplified approach which targets a proof of concept of the editor recommendation module rather than a fully functional one.

4.3.1 Complete Roles

The representative role distribution R_f of the featured class will be calculated by averaging over all articles under the *Featured* classification as given by the following equation;

$$\overrightarrow{R_f} = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{R_{f_i}}$$

; Where R_{f_i} is the role distribution of article i and n is the number of articles in the Featured class.

Given that an article was found to be of less than featured quality, missing roles needed to improve the article, $\overrightarrow{R_m}$ are computed in a straightforward manner using the following equation;

$$\overrightarrow{R_m} = \frac{1}{n}\overrightarrow{R_f} - \vec{R}$$

;where $\vec{R}_A$ is the original role distribution of the article, n is the ratio between the least role value in R and its corresponding value in $\overrightarrow{R_f}$. As such n acts a as normalizer of $\overrightarrow{R_f}$ as shown in the equation above, allowing is to extract the exact number of needed roles, by the ratios between $\overrightarrow{R_f}$ and $\vec{R}$.

The actual number of missing roles is determined by multiplying $\overrightarrow{R_m}$ by $|A|$, the number of editors who have already contributed to article A.

4.3.2 Sample Editors

Given the missing role values $|A|.\overrightarrow{R_m}$, and a role designated array of Wikipedia editors E_R, we create a sample array SR from array E_R containing $2|A|.\overrightarrow{R_m^i}$ editors for each role R_m^i in the missing role distribution $\overrightarrow{R_m}$.

By doing so, we end up with a sample of editors containing all the needed roles, from which we need to choose contributors to the article A that are expected to improve the quality of that article. Ideally, editor array E_R needs to be composed of all editors in the Wikipedia English community, however, for practical reasons; our editor array will only be composed of editors found in our sampled list of articles A_s.

4.3.3 Filter

Our reason for sampling twice the needed editor number, rather than not directly recommending editors to join the article editing process, is that, as proposed in section 4.2, roles on their own are not sufficient for deciding on the most appropriate editors for an article, but rather their relationship to an article, as reflected in their relationship network with this articles' editors, should play an essential role in mandating the usefulness of their contributions to the article.

As such in the filtering step we will choose half the editors in the role specific editor sample S_R collected in 4.2.3, this will be carried out by optimization of the editing network against the representative initiator matrix K_F of the *Featured* Class. This matrix is in turn calculated by averaging over all articles under the *Featured* articles classification as given by the following equation;

$$\Theta_F = \frac{1}{n}\sum_{i=1}^{n}\Theta_{F_i}$$

; Where Θ_{f_i} is the initiator matrix of article i and n is the number of articles in the *Featured* class.

The approach taken for that optimization can best be described by the following algorithm, which follows a straightforward hill climbing process with random restarts. Hill climbing simply means that at each moment a choice of editor is made from those possible based on the one that makes the Kronecker initiator of the editing network most similar to the Kronecker initiator of the *Featured* class. Random restarts means that occasionally the algoritm would return a chosen editor and reselect another one from the redundant sample S_R, which decreases the possibility of getting stuck in a local maximum. This algorithm will follow the steps below, which are also presented in the pseudo code in the next page.

1. Randomly iterate through all different roles in sample editors list.

2. Use hill climbing to decide which editor to choose under each role at each step by re-computing initiator matrix of A and comparing it to the representative initiator matrix of the Featured class

3. Repeat iterations until all needed roles are filled.

Algorithm 2. Hill Climbing Editor Selection

Input: Article A, Number of runs through roles array N, Number of role types T, Sampled editors by role SR, Initiator matrix of the Featured class Θ_F, Initiator matrix of article A Θ_A

-> Start
For i=1 to N
For j=1 to T
if($Length(SR_j) > 0$)
e <- $\arg\min_{SR_{j,k}} \| \Theta_F - \Theta_A \|$
Remove $\arg\min_{SR_{j,k}} \| \Theta_F - \Theta_A \|$ from SR_j
A <- A + e
Θ_A <- KronFit(A)
End
Return A

Chapter 5

Experimental Setup

To evaluate the three parts of our model, Role Clustering, Article Ranking and Editor Recommendation, we execute our algorithm, described in Chapter 4, on a sample of articles from the categories of *Social Sciences, Natural Sciences* and *Arts.* For better evaluation of the strength points of our complete algorithm, we evaluate our algorithm against several baselines. This Chapter explores implementation details of our algorithm including our data collection approach, the parameters we use for our algorithm, the baselines we created and our means for evaluating the quality of the algorithm's outcomes.

5.1 Data Collection & Use

The data collected for the thesis is divided into two segments, sample editor-centric data for generating role clusters and understanding general role behaviors in Wikipedia, and sample article-centric data for training the complete model. Given the extremely large size of data readily available from Wikipedia pertaining to articles and editor patterns, both types of data are sampled in a manner minimizing loss of generality as will be described below.

5.1.1 Role Clustering Data

Since our purpose is to construct natural role clusters adequately representing the different roles people assume in an article editing process, a large sample size was required to adequately represent the variations and nuances in the roles adopted by Wikipedia editors. From the 250+ million edits made to the English Version of

Wikipedia between January 2001 and August 2010 by 4+ million users, 20+ million edits were sampled representing the full editing records of 30000 editors in the Main, User and Wiki namespaces. The sampling was done with a bias for editors active in the previous two years, as it is expected that behavior in Wikipedia may have evolved from its time of inception, reflecting on the types and number of roles adopted by its editors.

5.1.2 Ranking Model Data

To train and test our ranking model, we take a more focused approach than that taken for collecting role cluster data, where we sample articles from 3 specific areas *Social Sciences*, *Natural Sciences* and *Arts*, the rationale behind that method of data collection is that quality of an article in relation to its editing network may be highly contingent upon the nature of that article, where one would expect articles discussing advanced technical subjects to require a different kind of editing relations from articles discussing biographies or general knowledge subjects. This is done to be able to test our model exclusively on each category to be able to discern the flexibility afforded by our model due to its content blind nature, where it can be trained and used on different types of collaboration outcomes independently, with limited underlying commonality between the generated models.

Furthermore, specific categories were chosen for sampling articles because one would expect the articles from a particular category to have a comparable number of edits and editors, which allows an adequate comparison of the editing networks of featured, regular and good articles in that category without being biased large editing size variations between these 3 classes.

Social Sciences, *Natural Sciences* and *Arts* were chosen in particular because they provide a sizeable number of *Good* and *Featured* articles unlike the rest of Wikipedia which currently contains fewer than 4000 featured articles, fewer than 12000 good articles and over 4 million regular articles, which allows the sampling to contain a comparable number of these 3 classes to be able to adequately map the differences between them.

Category	Featured	Good	Regular	Any Class	Category Total
Social Sciences					600
Business	20	20	15	55	
Economics	30	32	31	93	
History	75	73	77	225	
Politics	85	70	72	227	
Natural Sciences					600
Biology	80	80	80	240	
Health	80	80	80	240	
Physics	60	60	60	180	
Arts					600
Art	80	80	80	240	
Architecture	80	80	80	240	
Literature	60	55	65	180	
Class Total	**610**	**590**	**600**	**1800**	

Table 5.2. Detailed breakdown of sampled articles by category and classification

From *Social Sciences*, we sampled 210 Featured articles, 195 Good articles and 195 regular articles, amounting to a total of 600 articles Sub categories sampled include Economics, Business, History and Politics.

From *Natural Sciences*, we sampled 200 Featured articles, 200 Good articles and 200 regular articles, amounting to a total of 600 articles. Sub categories sampled include Biology, Health and Physics.

From *Arts*, we sampled 200 Featured articles, 195 Good articles and 205 regular articles, amounting to a total of 600 articles. Sub categories include Art, Architecture and Literature.

For each article in the sample, we collected its complete editor list including anonymous (IP) editors, which amounted to a total of 297,794 unique editors representing close to 10% of the full userspace of Wikipedia English,. These editors have a total of 200+ million edits, however, when taking into account Bot users who have not been properly designated as such, this number scales down to about 180+million edits. For data collection limitations, we only parsed the edit history of a user for up to a maximum of 5000 edits, leaving us with a sizeable 100

million edits to include in our model, which represents almost one-fourth the edits made to Wikipedia English until August 2011.

5.2 Baselines

To examine robustness and completeness of the proposed algorithm, it is compared against different baselines. The first baseline tests the structural metrics of the role detection part of our algorithm. The second baseline tests the role detection part of our algorithm. The third baseline tests the initiator matrix extraction part of our algorithm.

5.2.1 Stats Rank

This baseline provides a simplified means for ranking Wikipedia articles, by both ignoring the network of editors of the article and only clustering user roles through editing statistics of each user without considering the metrics derived by analyzing a user's network of articles. This baseline has two benefits, first it establishes means for evaluating usefulness of network metrics for creating user roles. Second, it allows us to establish the extent of complexity of the issue we are trying to resolve by showing how an intuitive straightforward means of evaluation can perform the ranking task without any structural understanding of the article and editor networks.

5.2.2 Role Rank

This baseline is built on using only the distribution of roles in an article, given by the percentage of editors adopting each role and weighted by the number of edits made by each editor, for learning a ranking for the quality of that article, where the ranking process in that case will be simply to suggest what kinds of roles need to be more prevalent in the article, either by recommending new editors doing these roles or by increasing the contributions of existing editors with these roles. This baseline provides a benchmark against which effectiveness of the full *kronecker-based* approach may be evaluated.

5.2.3 KronFit Rank

This baseline is a variation of the ranking algorithm which does not include the process of role detection in the article ranking/improvement process. As such, the article editing network is uni-modal where all nodes are considered to be editors with no added complexity of node type. To perform the ranking algorithm, the ranking model is constructed solely by carrying out the KronFit process as formerly described without including role distributions as features in the classification process. In addition to its usefulness for testing leverage offered by adding the role detection process, this method also presents a simpler approach to augmenting collaborative environments requiring limited content specific feature construction, as is the case in the role detection part.

5.3 Article Ranking

Strictly automated evaluation metrics are used for testing performance of the complete article ranking model in relation to the different baselines discussed in 4.2. These metrics include precision and recall. These metrics will be incubated in a cross validation framework of the ranking model. Section 5.3.1 further explains what Cross validation entails.

5.3.1 Cross Validation

N-fold cross validation is a commonly used and simple method for testing predictive performance level of a model developed on new data, by breaking down the training data into N partitions, training a model on N-1 partitions and testing it on the remaining partition, where each partition is used once for testing, providing N models and performance scores. By averaging these scores, as given by the equation below, a predicted performance of the model on new training data can be established.

$$Performance = \frac{1}{N}\sum_{P} Performance\ on\ Partition\ P \text{ ; where P=\{1.....N)}$$

For our testing purposes, we will use 10-fold cross validation which provides a compromise between the accurate results obtained by large values of N and the computational efficiency of small values of N. We use the Rapid Miner analysis environment to perform the cross validation process and calculate precision and recall metrics for the classification component of our algorithm

5.3.2 Category Specific Evaluation

Since our model rests mainly on content independent analysis of collaboration patterns, our algorithm needs to be tested on different categories individually. Doing so allows us to confirm the model's adaptive ability, which is the cornerstone of our thesis, as by being able to adapt effectively, our algorithm can then be used for augmenting different areas of collaboration, where the content requires a different skill set than that needed for editing articles. Therefore, using each of the 3 category samples described earlier, we will execute different runs of our algorithm, and evaluate the results in the same manner carried out for the whole sample collectively.

5.4 Editor Recommendation

For adequate evaluation of our complete paradigm, we need to properly measure the proof of concept approach to recommending editors presented in 4.3. That being said, evaluation of the editor recommendation stage is not as straight-forward as that of the article ranking stage, due to the fact that, if we were to recommend editors to join regular articles in hopes of improving its rankings, we would need to actually run that experiment and wait for the ranking change of the article.

To overcome that obstacle, we create one simple modification, instead of recommending editors for existing *Non-Featured* articles in the sample; we deliberately distort *Featured* article editing networks instead, particularly those closer to the representative network parameters of a *Featured* article. This distortion is achieved by removing 20% of editors of a *Featured* article A in a random fashion. Following that, we perform the editor recommendation algorithm

described in 4.3, with the minor modification of using a sample editor list SR containing the 20% removed editors and an additional randomly selected equal number of editors from the 297,794 editors collected in our second dataset with the same roles as those designated by the originally removed editors but with no direct relationship with article A.

By adopting that approach we can easily test whether our editor recommendation model offers any leverage in assisting collaborative networks by comparing the percentage of originally removed editors from A who were recommended for rejoining the network to the percentage likely to be obtained by random selection of editors from SR, which naturally tends towards 50% by virtue of the symmetry created by making SR contain half real members of the network and half fake members.

Chapter 6

Results

"However beautiful the strategy, you should occasionally look at the results."

6.1 Role Modeling

Applying the *K-Sampling* algorithm described in Chapter 4 using statistical and structural properties listed in 4.1, we obtained the results presented in figure 6.1. The figure depicts a centroid plot, where each centroid is designated by one of 5 colors. The x-axis contains the different features of each centroid and the y-axis gives the value of each feature for a centroid.

Contrary to our initial guesses, **(M2)** Article Creation Size, **(M6)** Contribution Size and **(M10)** minor edit ratio proved little relevance to detection of editor roles, giving nearly identical values for all 5 centroids. The following sub section analyzes the centroids generated through the K-Sampling algorithm and identifies *Spontaneous Editors, Opinion Promoters, Collaboration Coordinators, Question Marks* and *Topic Contributors* as the 5 key roles designated by the centroids of the clustering model.

6.1.1 Analyzing Centroids

Centroid 0: ***(Designated by the dark blue line)*** This centroid can be distinguished by very high article and category distributions (M3,M5) and a very low clustering coefficient (M13). All signifying that members of that centroid edit relatively unrelated articles. When coupling the low clustering coefficient with slightly below average connectivity, and low project and talk ratios, in addition to being the largest cluster in terms of sample members, this centroid can be considered to reflect editors who move from one article to another in a spontaneous fashion making edits along the way, fitting the role of a *Spontaneous Contributor*.

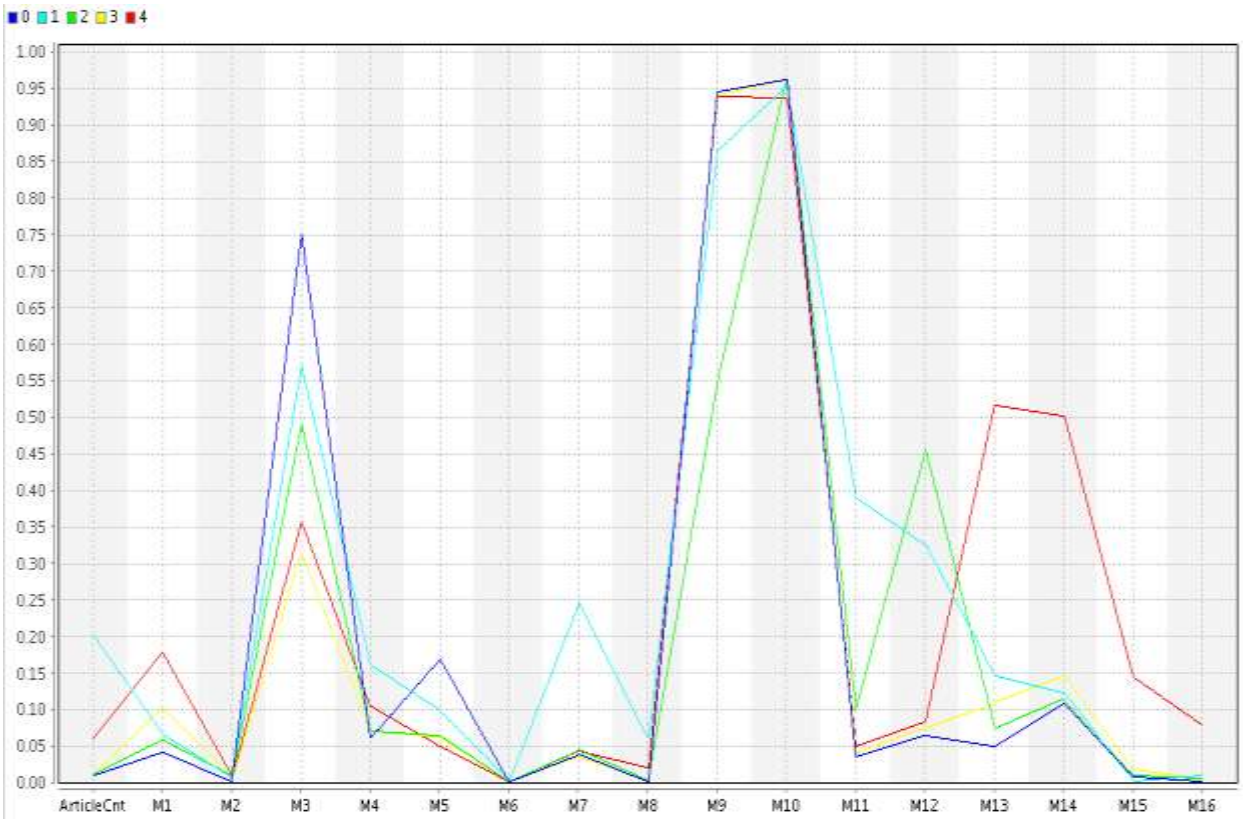


Figure 6.1. Role Clustering Centroids Plot

Centroid 1: ***(Designated by the light blue line)*** Typical members of that role normally contribute to already developed articles (high average edit time **M4**), they may tend to get involved in a lot of opinion feuds (high revert ratio **M7**), edit interrelated topics in a distributed fashion (average clustering coefficient **M13**) and actively contribute to projects and article discussions(high talk and project ratios **M11** and **M12**). In addition, this cluster has the smallest number of members, constituting less than 10% of the editor sample. Despite being hard to label, this

mixture of characteristics fit well with the role of an *Opinion Promoter*.

Centroid 2: ***(Designated by the green line)*** Typical members of that centroid spend most of their time in discussion pages as they have the highest talk ratio **(M12),** and are active in Wikipedia projects, as mandated by their above average project ratio. This description fits well with the role of a *Collaboration Coordinator.*

Centroid 3: ***(Designated by the yellow line)*** Aside from having a low article distribution, that centroid lacks any other distinguishing factor and most of its metrics falls within average values. This could either hint at the need for varying the number K of clusters used in the model, or reflect a class of Wikipedia editors who do not fall under a stable kind of role. Given that the number of members of that centroid exceeds 30% of the sampled users, we are inclined towards the second justification. That being said, editors falling under that centroid will simply be labeled as a *Question Mark.*

Centroid 4*: (Designated by the red line)* Typical members of that centroid create the most articles (high article count **M1**), focus on a few articles (low article and category distributions **,M3 and M5)** and these few articles are part of very specific topics, as mandated by having the highest structural metrics **(M13,M14,M15,M16).** These characteristics fit well with the role of a *Topic Contributor*.

Aside from the issue of the *Question Mark* centroid, results from the aforementioned role clustering process agree well with our top down approach for identifying key roles in the Wikipedia network, specifically, Centroids 0, 1 and 4 fit with our conception of *Spontaneous, Distributed* and *Topic Contributors* presented in section 4.1.

6.1.2 Role Distributions

In an attempt to better gauge the correctness of our role analysis and to understand how these different roles influence the quality of an article, we used the 5 aforementioned role centroids to construct representative role distributions of *Featured, Good* and *Regular* Article classes. These representative distributions were created by averaging over the distributions of articles in these respective

classes as given by the following equation.

$$R_{C_{rep}} = \frac{1}{n}\sum_{i=1}^{n} R_{C_i}$$

; Where R_{c_i} is the role distribution of an article i in class C, and n is the number of articles in class C.

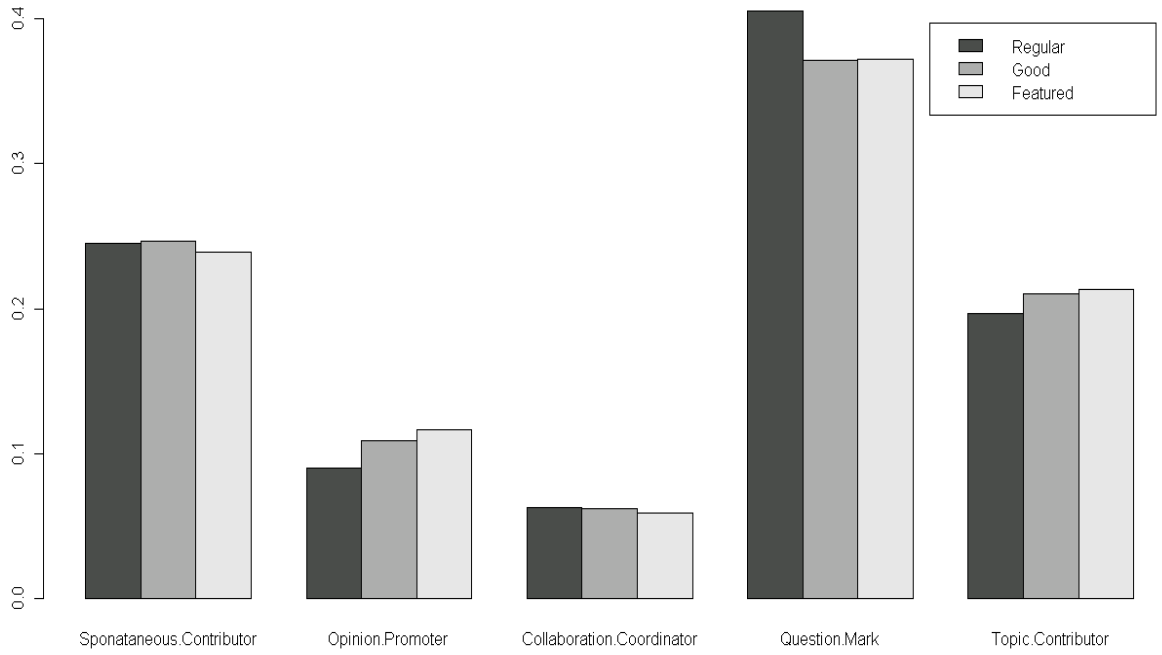


Figure 6.2. A comparison of representative role distributions of *Featured, Good* and *Regular* articles

The 3 generated distributions are represented on the bar plot in figure 6.2 above. As mentioned in 2.2, article comprehensiveness is a key criterion of *Featured* articles. As such, a balanced increase in the portion of *Opinion Promoters* in the article editing process would be expected to boost that needed comprehensiveness by allowing the interchange, dialogue and sometimes conflict of varying opinions and perspectives, provided that it is balanced by an equal increase in *Topic Contributors*, who can better promote accuracy, verifiability and

proper articulation of information. Such paired increase in ratios of *Topic Contributors* and *Opinion Promoters* as we move up article classes from *Regular* to *Featured* Articles is evident with consistency in the bar plot, where the ratio of *Topic Contributors* increased by 7% from 20% in Regular articles to 21.4% in *Featured Articles* and the ratio of Opinion Promoters increased by over 25% from 9.0% in *Regular* Articles to 11.7% in *Featured* Articles.

Moreover, the plot shows a slight decrease in the ratio of *Collaboration Coordinators* as we move towards *Featured* Articles. This could be interpreted in three varying ways, either that kind of role is not particularly needed in editing articles as such it is more prevalent in regular articles, or it is only important for lower quality articles which needs more coordination for enriching its content, or possibly that variation, given its limited extent, is due to biases of our sample and does not incur a concrete conclusion.

As for *Spontaneous Contributors* and *Question Marks*, while they generally decrease in ratio from *Regular* to *Featured* articles, by virtue of the increase of other roles, however, that decrease does not show a consistent pattern and possibly implies our need to consider more features for reformulating these two roles.

6.2 Article Ranking

The results of article ranking using the complete ranking model in addition to the *StatsRank, RoleRank* and *KronFit* baselines applied to the 1800 article samples is presented in table 6.1 The reported results are those relating to the recall and precision of the *Featured* article classification. Confirming our initial guesses, the results show a leverage offered by incorporating the *Role Distributions* and *KronFit* components in parallel to produce article classifications. However, the difference between StatsRank and RoleRank baselines' performance is not significant, which either suggests the absence of a particular relationship pattern between contribution networks of users and their roles, or the need to include more sophisticated metrics in that process.

Baselines	Recall	Precision
StatsRank	0.64	0.701
RoleRank	0.66	0.691
KronFit	0.6	0.642
Complete Model	0.77	0.83

Table 6.1 Cross Validation performance of the complete model and the different baselines applied to 1800 articles evenly sampled from *Social Sciences, Natural Sciences and Arts* categories

Furthermore, results of independent application on each of the 3 categories are presented in table 6.2. Interestingly enough, results obtained show significant differences in the performance values for each of the 3 categories, with performance on social sciences being most promising. The reason for this may be driven by one of several possibilities. First, the sample size may not be large enough to establish statistical significance of the performance differences and the large discrepancy between the number of articles in Wikipedia and the number of sampled articles may have induced different performance biases for the 3 categories. Second, different categories could require different patterns of collaboration and our algorithm might be best fit to handle some of these patterns more than others. The next section will help further explore that issue providing samples of various networks.

Baselines	Recall	Precision
Social Sciences		
StatsRank	0.68	0.71
RoleRank	0.71	0.73
KronFit	0.64	0.68
Complete Model	0.84	0.86

Natural Sciences		
StatsRank	0.653	0.664
RoleRank	0.671	0.69
KronFit	0.63	0.672
Complete Model	0.79	0.81
Arts		
StatsRank	0.62	0.635
RoleRank	0.61	0.645
KronFit	0.61	0.646
Complete Model	0.73	0.768

Table 6.2. Cross Validation performance of the complete model and the different baselines by independent application on Social Sciences, Natural Sciences and Arts categories.

6.2.1 Network Samples

This section explores key initiator matrices of the *Featured* and *Non-Featured* article classes, and presents articles from both classes which illustrate the patterns expressed by these initiators.

Interpreting an initiator matrix K_0

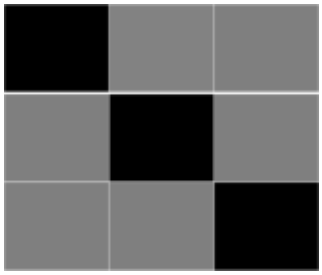


Figure 6.3. Initiator Matrix Illustration

As *KronFit* follows a bottom up approach, by which it generates a mapping from real networks to the compressed format of an initiator matrix K_0, it may sometimes be challenging to directly understand qualitative properties of the real network from the parameters of K_0. However, as emphasized in Leskovec et. al. (2010 a),

generally, for self-edges of an initiator network, designated by the black diagonals in figure 6.3, higher values tend to indicate an increase in the degree of *homophily*(See section 3.1) of the network. If more than one of these diagonal values is high, it can be interpreted as an increase in the number of macro-communities in the network. Higher values of the cross-edges, *designated by the grey locations,* can indicate an increase in the degree of *heterophily*(See section 3.1) of the network. Moreover, generally high values of the parameters in their totality imply a denser network.

Key Initiator Networks

Since we extracted for each article a 3x3 initiator matrix, a single representative initiator matrix for any class of articles could potentially be constructed by the average of the initiator matrices of the articles in that class obtained through the following equation;

$$K_{C_{rep}} = \frac{1}{n}\sum_{i=1}^{n} K_{C_i}$$

; Where K_{C_i} is the Kronecker initiator of an article i in class C, and n is the number of articles in class C.

Since our purpose is both to understand the common properties of *Editor Networks* of the *Featured* and *Non-Featured* articles classes and to understand the differences in properties between both classes, therefore, along with the initiator matrices presented for the *Featured* and *Non-Featured* classes, we constructed a modified matrix $K_{F/NF}$ by dividing every location in K_F, the initiator matrix of the *Featured* articles class by every corresponding location in K_{NF} the initiator matrix of the *Non-Featured* articles class. Figure 6.4 below presents all the three aforementioned matrices visualized as a Heat-map, where increased redness of a matrix location implies a higher value of that location.

In reference to figure 6.4, note the general similarity of the properties of (A) K_{NF} and (B) K_F. The pronounced high intensity of the top left self-edge (0, 0) in

both matrices implies that Wikipedia editor networks generally tend toward homophily, where similar degree editors tend to be connected. Moreover, the varying intensities of the 3 self-edges (0, 0), (1, 1) and (2, 2) in (A) and (B) implies that generally editor networks tend to cluster in one large community with possibly few smaller communities. The fact that the three highest edge intensities in both matrices (0,0) , (1,0) and (0,1) are in the top left corner implies that within larger communities editors tend to exhibit more heterophily by being connected to editors of different degree ranges.

As for the comparative initiator network (C) $K_{F/NF}$, the increasing intensity of self-edges as we move from location (0,0) to (2,2) means that the percentage differences of intensities of the 3 self-edges in (B) K_F is less than that in (A) K_{NF} , which implies that communities in *Featured* articles tend to have more similar sizes than those of *Non-Featured* articles. As for pronounced intensities of most cross-edges(off-diagonal locations) in $K_{F/NF}$, they imply that *Featured* articles generally tend to have more connectivity and heterophily than *Non-Featured* articles.

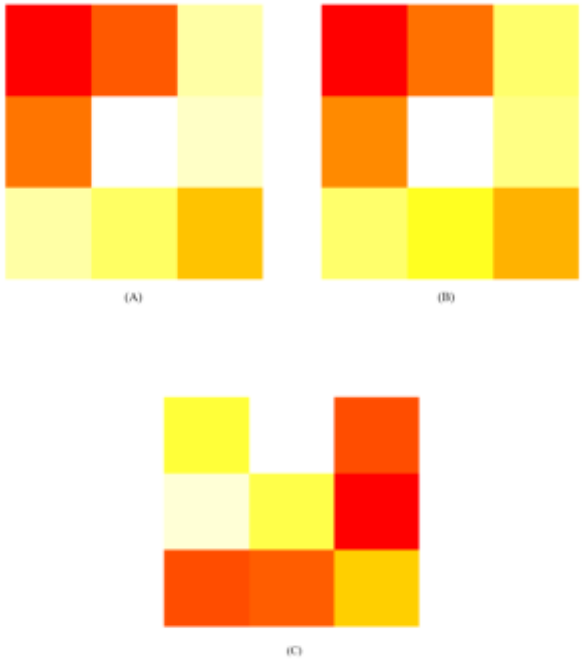


Figure 6.4: (A) Non-Featured Initiator (K_{NF}) HeatMap.(B) Featured Initiator (K_F) HeatMap. (C) Comparative Initiator HeatMap ($K_{F/NF}$)

To test our aforementioned analysis, we extracted the four articles most representative of *Featured* and *Non-Featured* classes, which are the four articles whose initiator matrices are closest to their classes' representative initiator matrices as given by the following equation;

$$\{A_{C_1}, A_{C_2}\} = \arg\min_{A_i, A_j} \| K_{C_{rep}} - K_{A_{i|j}} \|$$

; Where $K_{C_{rep}} = K_F$ if C=*Featured* and $K_{C_{rep}} = K_{NF}$ if C=*Non-Featured*

The four resulting articles, *Lindow Man and Las Meninas* for the *Featured* Class, *and Public Art and Conflict Theory* for the *Non-Featured* Class are presented in Figures 6.5 and 6.6 respectively. *Featured* article editor networks are distinguished with a *dark blue* node color, whereas *Non-Featured* article editor networks are distinguished with a *light blue* color. Notice that role assignments were not included in these networks, as the purpose here is to correlate an initiator matrix's parameters to the properties of its original network.

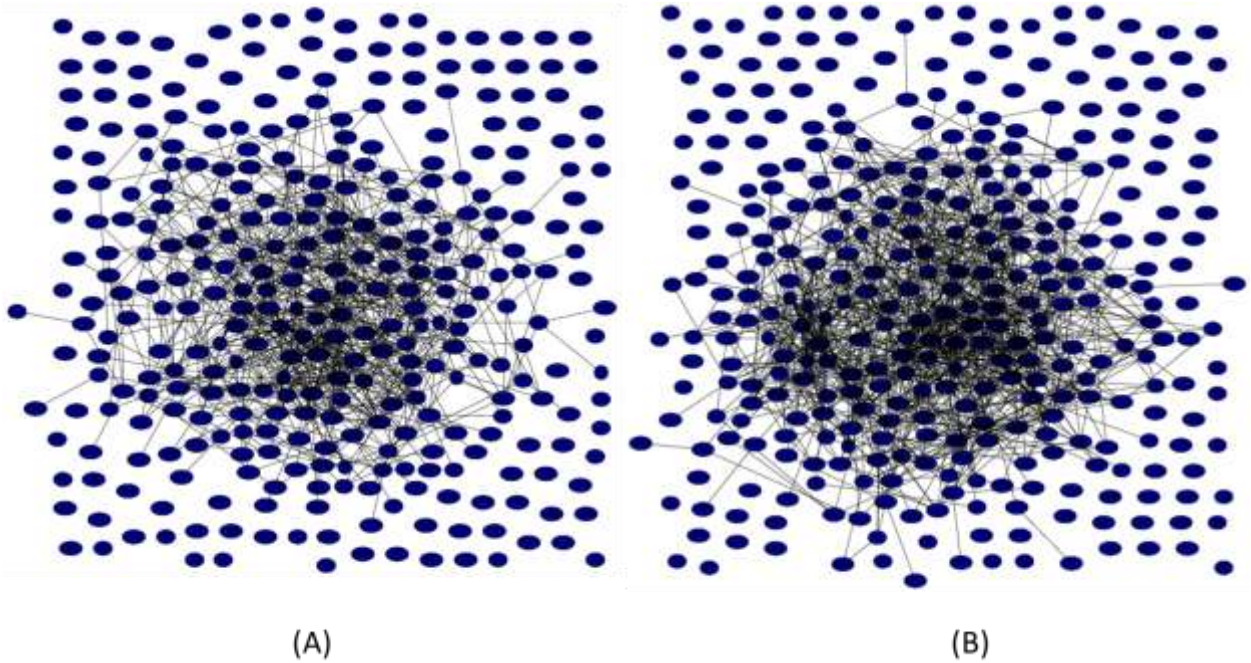


Figure 6.5. Top 2 *Featured* Class Networks (A) *Lindow Man* Editor Network. (B) *Las Meninas* Editor Network

As related to the 2 *Featured* articles in figure 6.5 above, notice the existence of one large central community encompassing almost all connected editors.

Furthermore, notice the general tendency towards *homophily*, where the high degree editors at the core tend to be connected to each other and lower degree editors at the periphery tend to connect as well. Despite these *homophily* tendencies, many connections can be observed from high degree editors at the core to lower degree editors at the peripheries by virtue of the generally high connectivity of *Featured* article networks hypothesized through the former analysis of the comparative initiator matrix $K_{F/NF}$

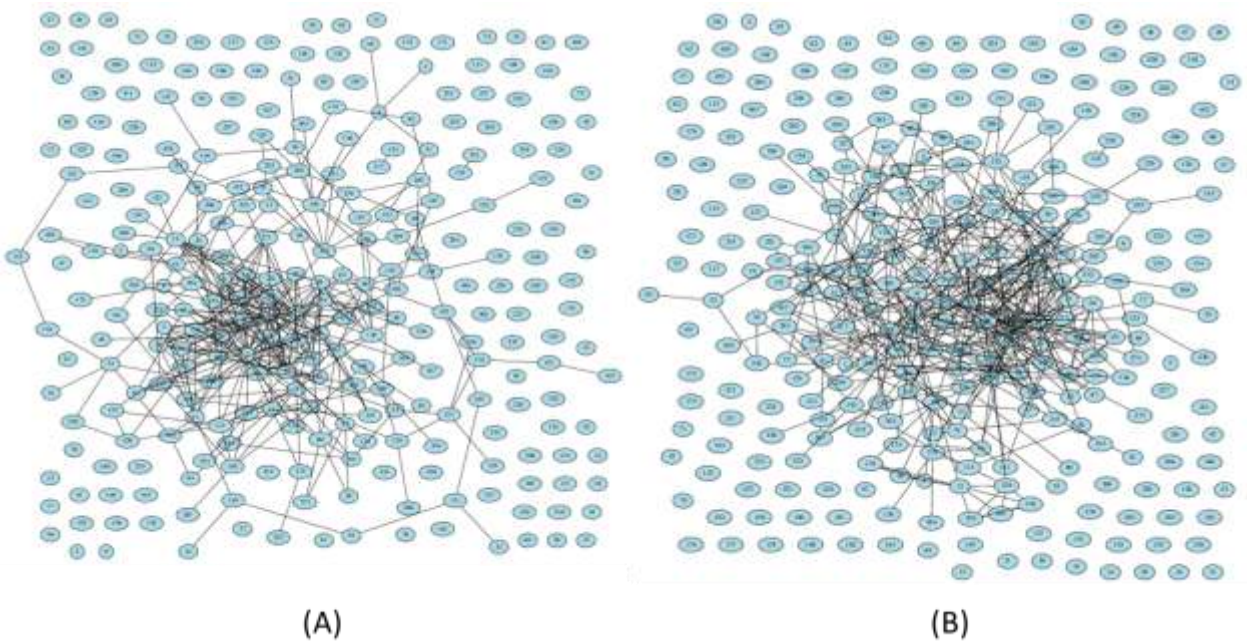


Figure 6.6. Top 2 *Non-Featured* Class Networks (A) *Public Art* Editor Network. (B) *Conflict Theory* Editor Network

As for the 2 *Non-Featured* articles in figure 6.6 above, notice how all connected editors form one large unbalanced community like hypothesized in analysis, which contains sub-community divisions as shown in lower left of (A) and the right half of (B), respectively. Furthermore, notice the high *homophily* of edge connections, especially in (*A) Public Art*, where the network can clearly be divided into layers of lower degree connected editors as we move away from the center of the network, and where very few connections exist between high degree connected editors at the core and lower degree connected editors at the peripheries, as inferred from our analysis of the comparative initiator matrix $K_{F/NF}$. The aforementioned pronounced *homophily* and community structure agree largely

with our former description of *Non-Featured* networks using its initiator matrix K_{NF}, which reinforces the high expressiveness of the very compressed mapping of an initiator matrix and its ability to adequately imprint the key properties of its original network.

6.3 Editor Recommendation

As emphasized in the methodology section, given the complexity of the task of completing editor networks of Non-Featured articles by network centric recommendation, this module of our complete algorithm was implemented as a proof of concept by following the simple hill climbing approach described in 4.3, in lieu of a potentially more sophisticated approach building on *Kronecker Graphs*. This section presents the performance results of that approach obtained by applying our algorithm to a sample of *Featured* articles in the manner described in section 5.4.

	Precision	Null Model Precision
Recommendation	0.80	0.5

Table 6.3 Performance of Editor Recommendation Module

As evident from the above table, the editor recommendation algorithm was able to extract the actual missing editors of an article with high precision; however, given the relatively simplistic means for testing the recommendation model, possibly real life applications will require more sophisticated network completion algorithms which build directly on the properties embedded within *Kronecker Graphs*.

Moreover, figure 6.7 below provides a sample run of the algorithm on the Lindow Man article used before in the article ranking samples. By observing (C) the recommended network, one can easily see the strong ability of the hill climbing approach in picking up high degree editors of the network from the sample set SR,

and it also becomes obvious that a major portion of the errors of the model are caused by nodes with 0 connections in the article which are real ly hard to discern or distinguish as being invalid as most editing network have been demonstrated to contain a good portion of isolated nodes as is the case here.

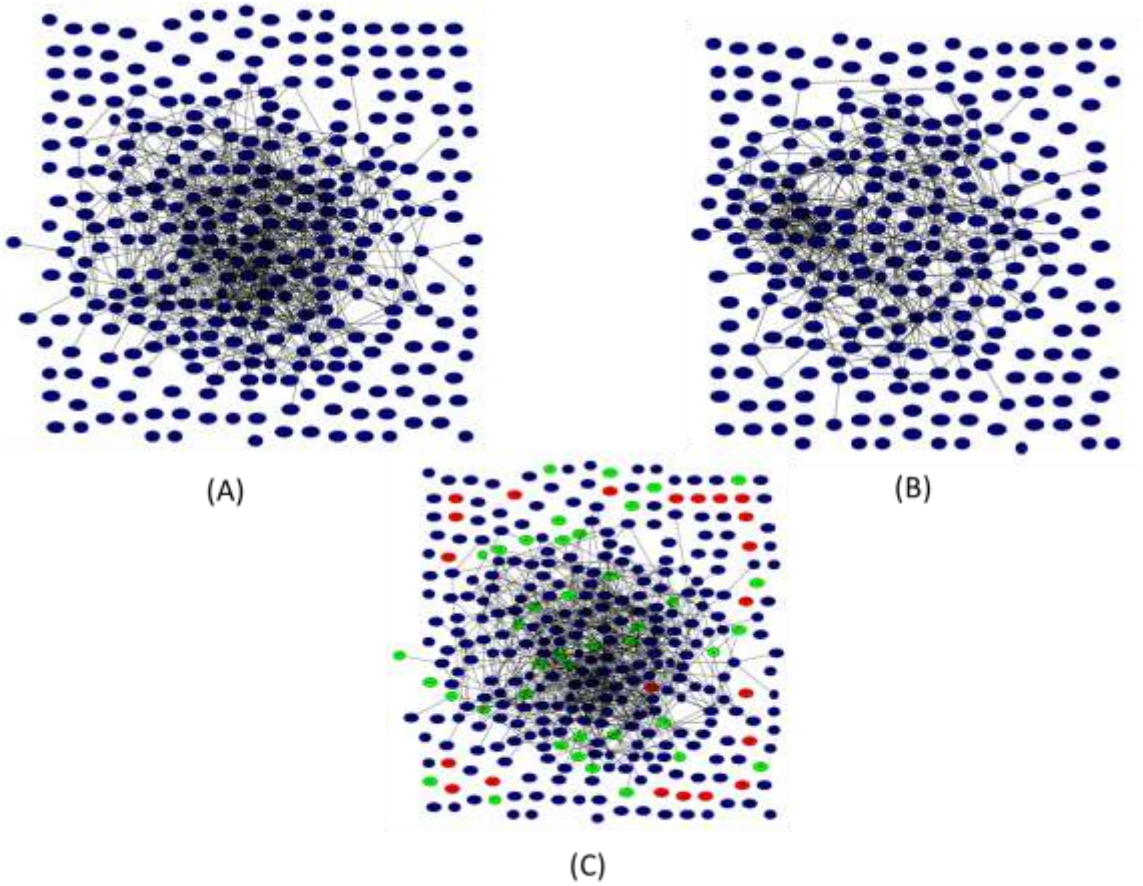


Figure 6.7 .(A) *Lindow Man* Editor Network. (B) *Lindow Man* Distorted Network. (C) *Lindow Man* Recommended Network

Chapter 7

Conclusions and Future Work

This thesis presented a paradigm for computational mediation of human- centered collaboration activities applied to Wikipedia articles. It relies on a three part framework which consists of Role Modeling, Article Ranking and Editor Recommendation. First, role types are generated by clustering a sample of Wikipedia editors. Then, an article is ranked using structure of its role designated editing network. Lastly, new editors are recommended for the article based on its ranking. This paradigm shows promising results as a stepping stone to automatically assisting human collaboration activities in a content independent manner.

Evaluation for both the article ranking and editor recommendation applications shows that our methodology strongly correlates with human rankings and content improvement patterns in Wikipedia. Exclusive application of the role distribution and *KronFit* components has manifested the leverage offered by combining both techniques in ranking and influencing articles.

We believe our work is open to numerous future research directions and applications, as generally related to large scale group collaboration, like that presented by Wikipedia. First, our ranking algorithm exhibits a relatively consistent accuracy for the collaborative environment of Wikipedia. With other more complex collaboration environments, where, for example, quality of collaboration outcomes is not directly provided like the case with Wikipedia, new approaches will need to be explored to establish quality of collaboration outcomes.

As related to the feature based combination of role distributions and the KronFit approach for ranking collaboration content, a more sophisticated approach needs to be researched which is able to learn key parameters of a collaboration network in a manner that is able to embed the different roles of collaborators directly in its representation rather than integrating them as separate features in the manner we provided in this thesis.

Furthermore, analyzing collaboration outcome can be extended from focussing on a single unit of collaboration, an article in the case of Wikipedia, to assessing an array of interrelated collaboration units and how to balance the improvement of a single unit with the improvement of the whole units using the same resources. This translates to Wikipedia in how to use the same number of editors in balancing the improvement of single articles and topics composed of a group of articles.

Finally, while the editor recommendation component manifested the ability to improve article content by recommending a certain group of editors, a potentially more holistic editor recommendation paradigm is one which directly employs Kronecker parameters to “complete” the editor network, like that adopted for the network completion problem in [21] , rather than simply using it as a fitness function for filtering editors.

Chapter 8

Appendix

This appendix provides a sample of key Featured and Non-Featured article editing networks.

8.1.1 Featured Article Networks

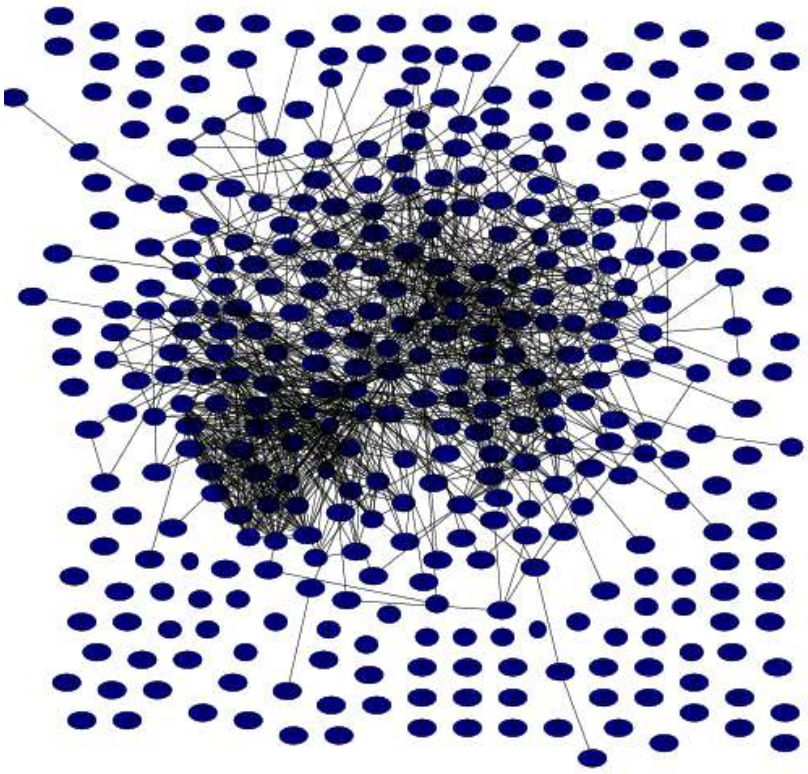


Figure 8.1. Featured Article: *Keratoconus* Editing Network

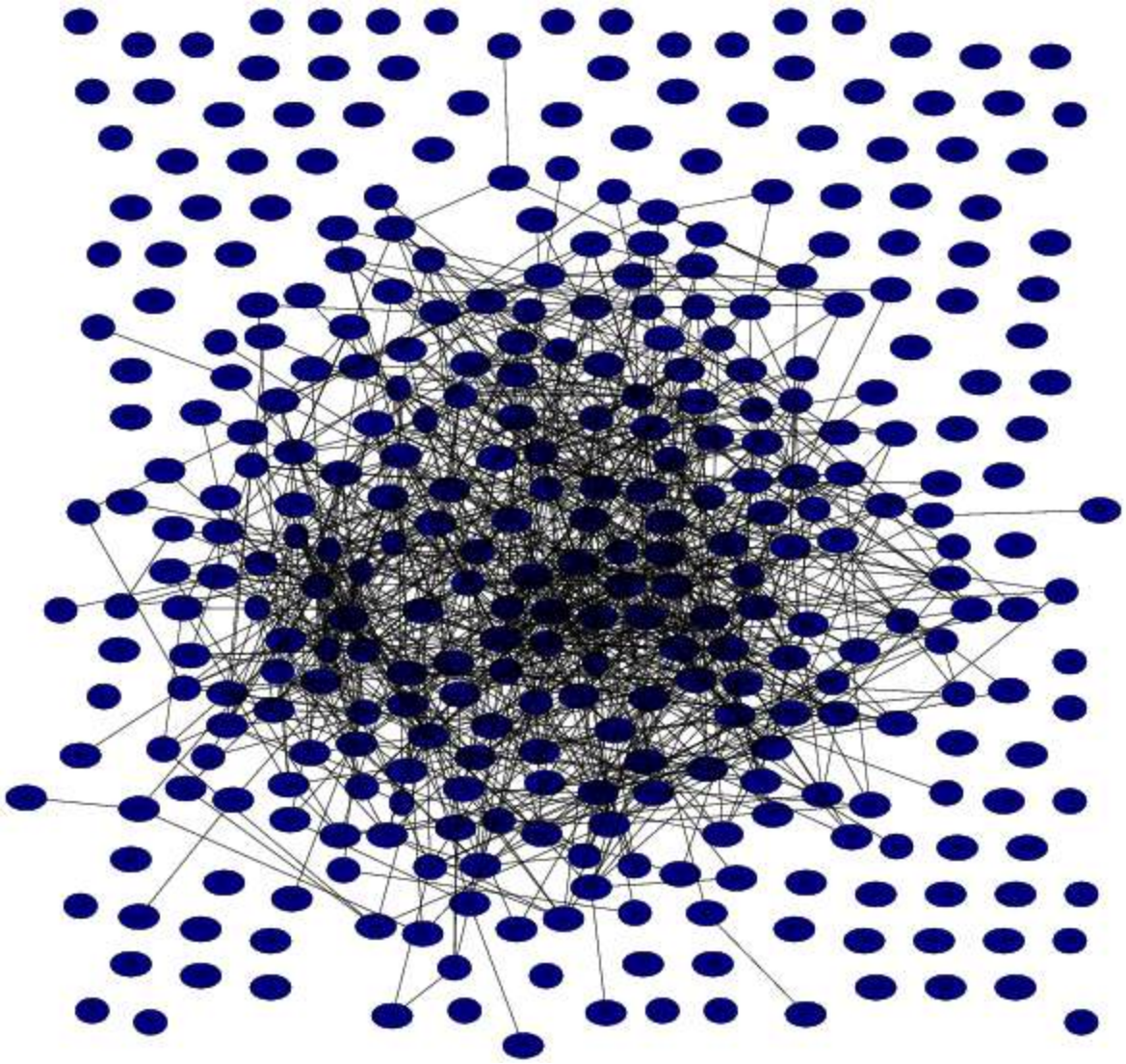


Figure 8.2. Featured Article: *Las Meninas* Editing Network

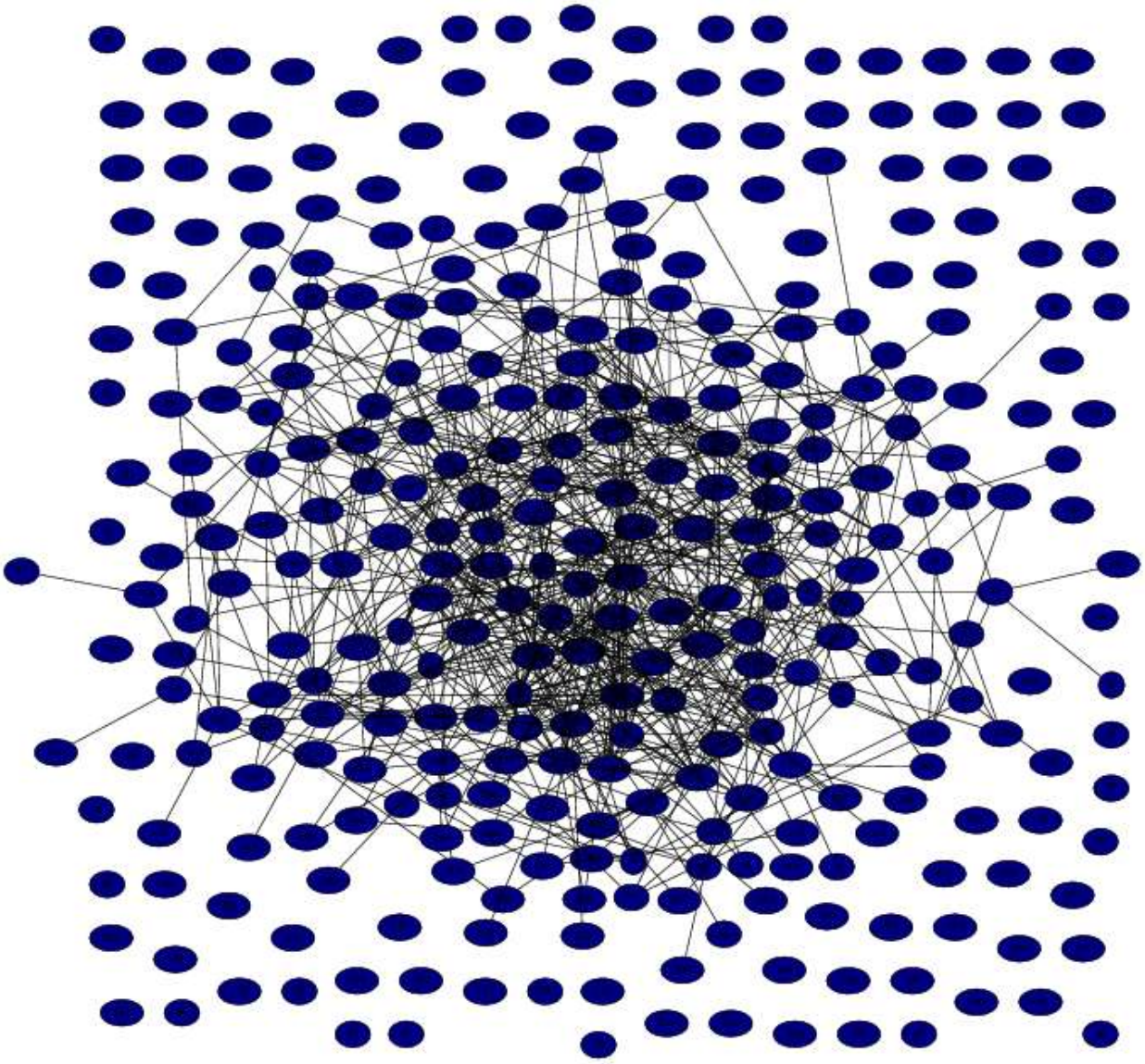


Figure 8.3. Featured Article: *Lindow Man* Editing Network

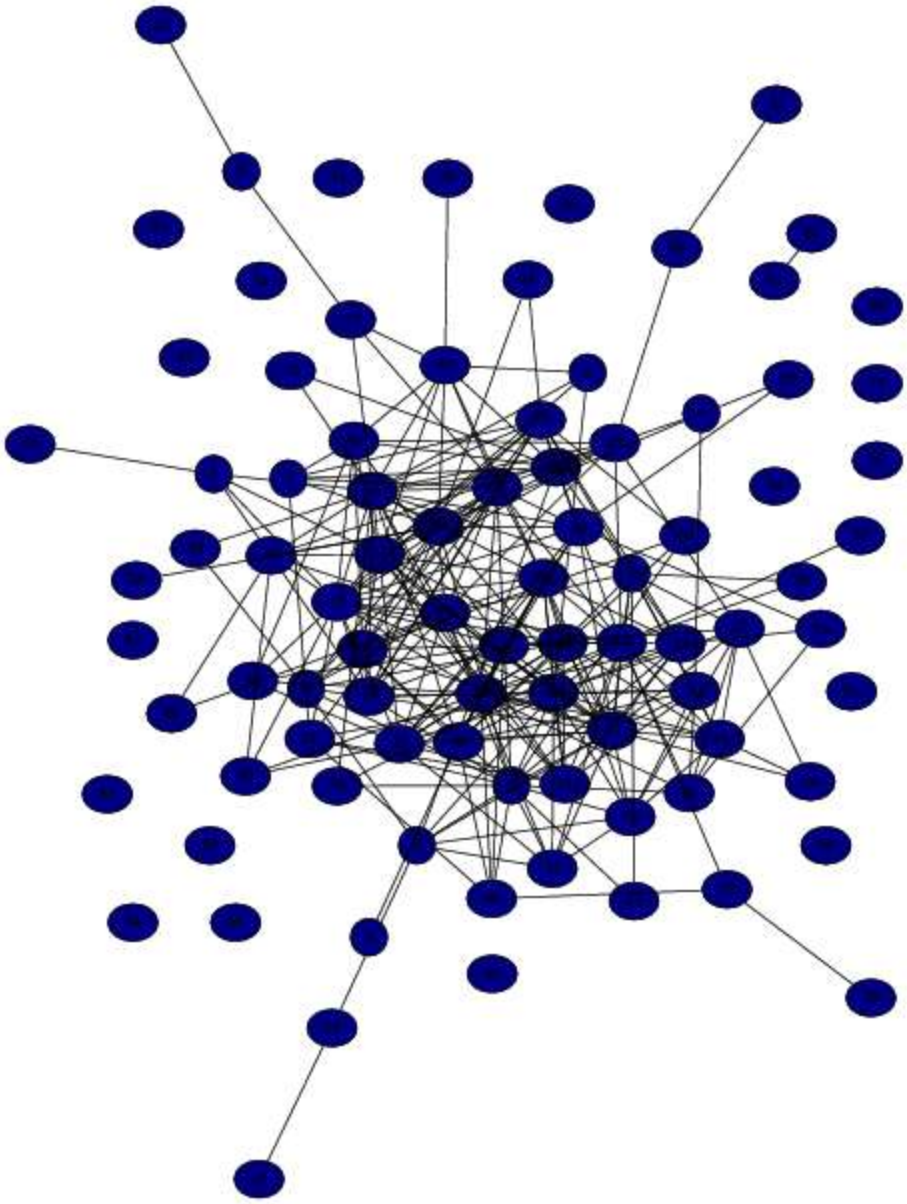


Figure 8.4. Featured Article: *New York's 27th Congressional district special election* Editing Network

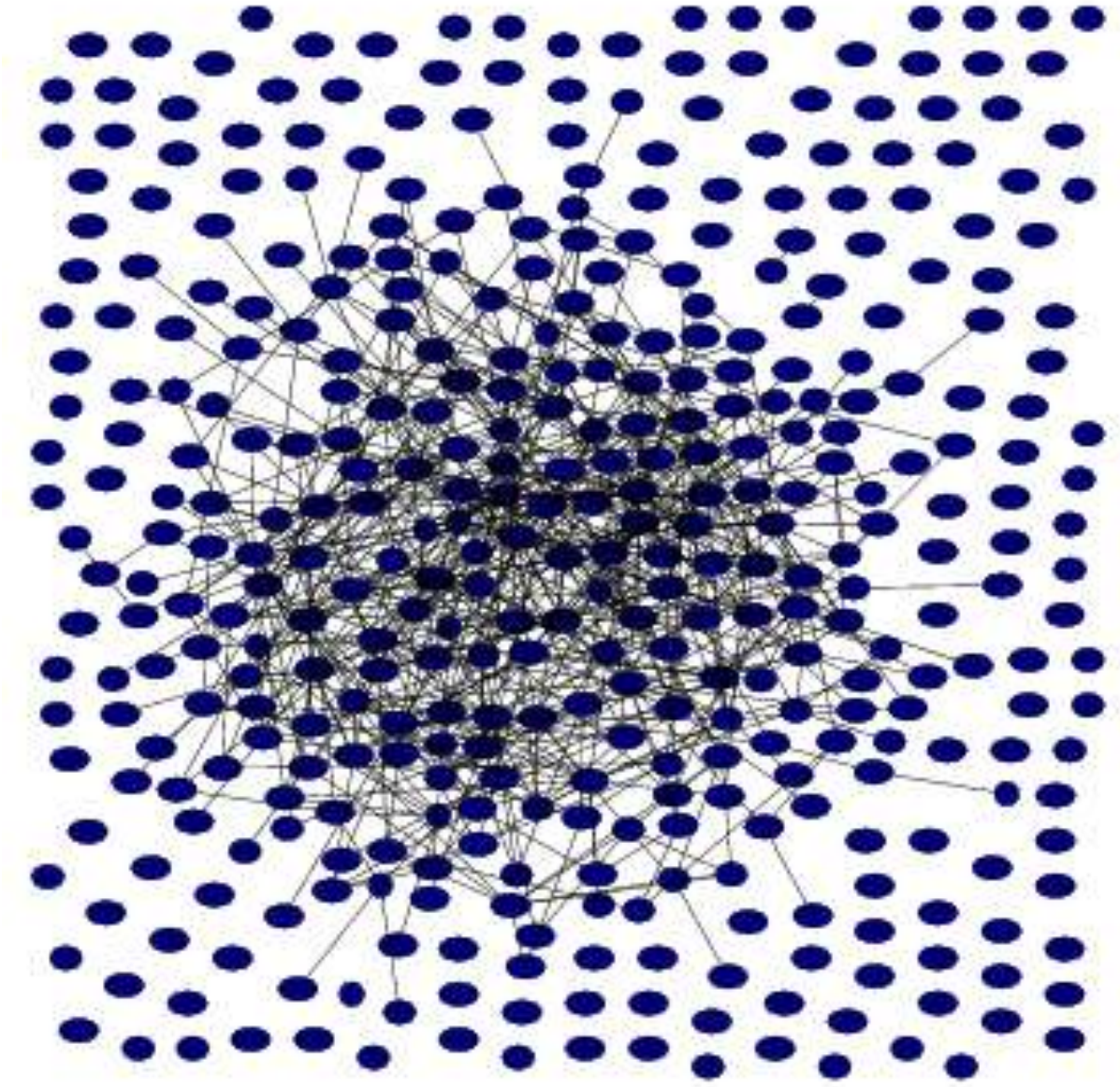


Figure 8.5. Featured Article: *The Time Traveler's Wife* Editing Network

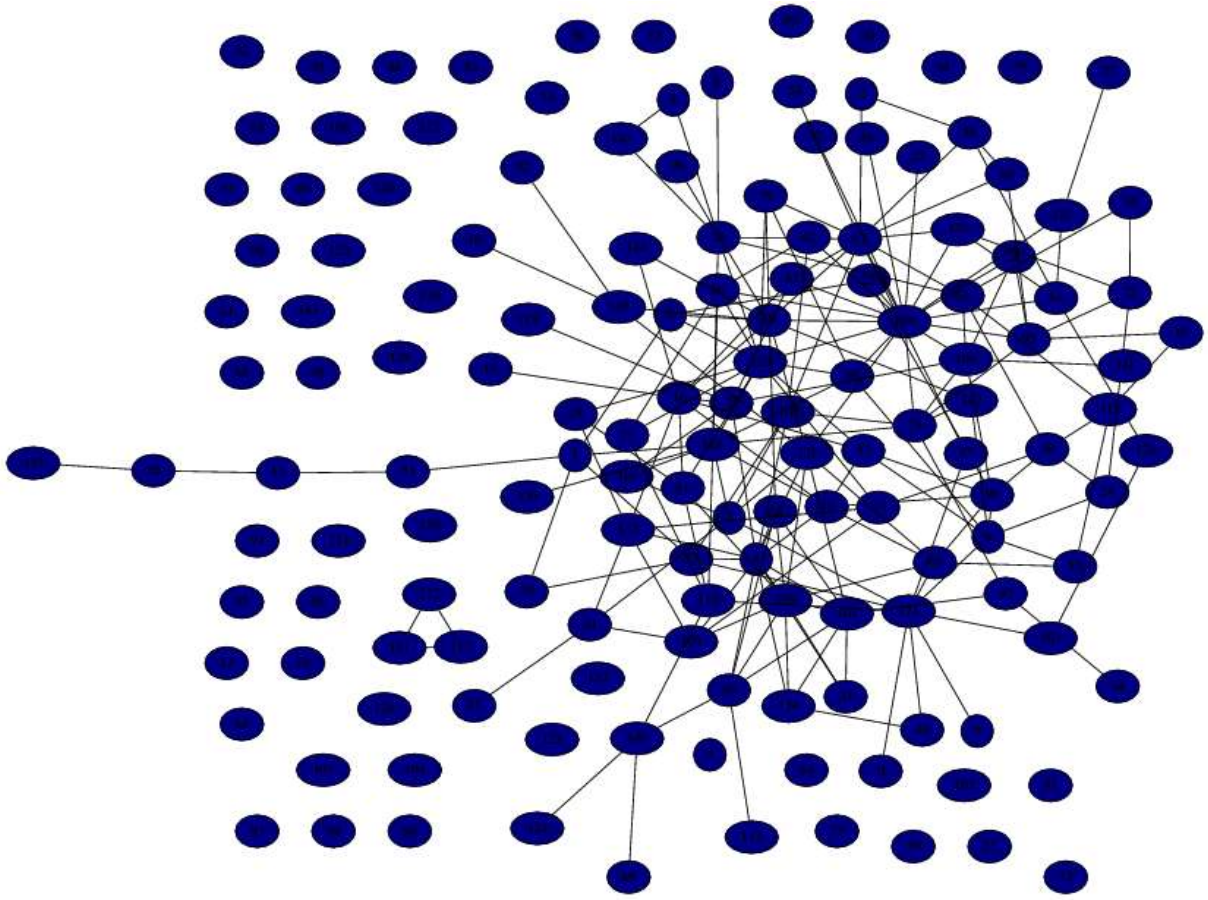


Figure 8.6 Featured Article: *The Country Wife* Editing Network

8.2 Non-Featured Article Networks

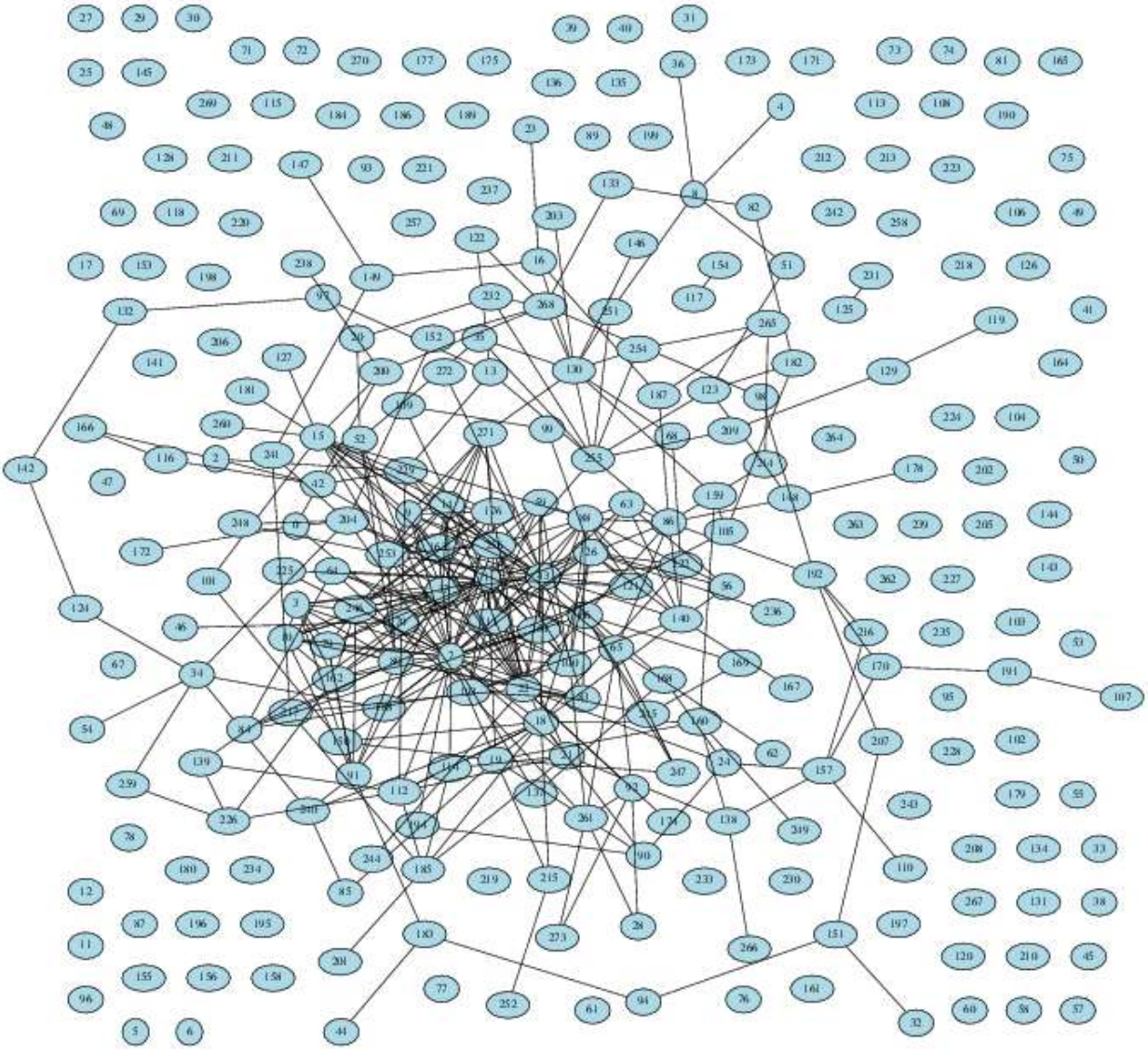


Figure 8.7 Non-Featured Article: *Public Art* Editing Network

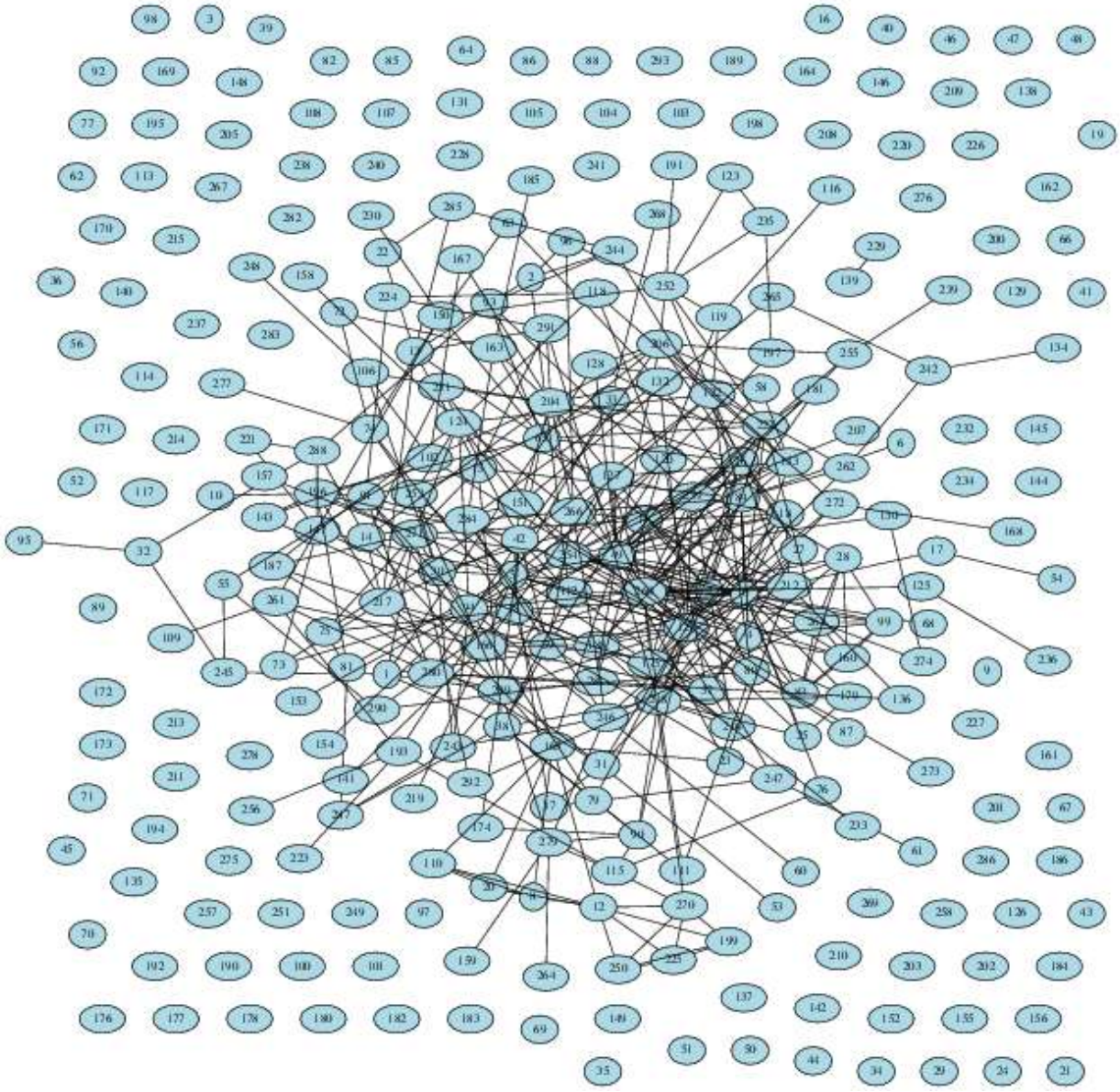


Figure 8.8 Non-Featured Article: *Conflict Theory* Editing Network

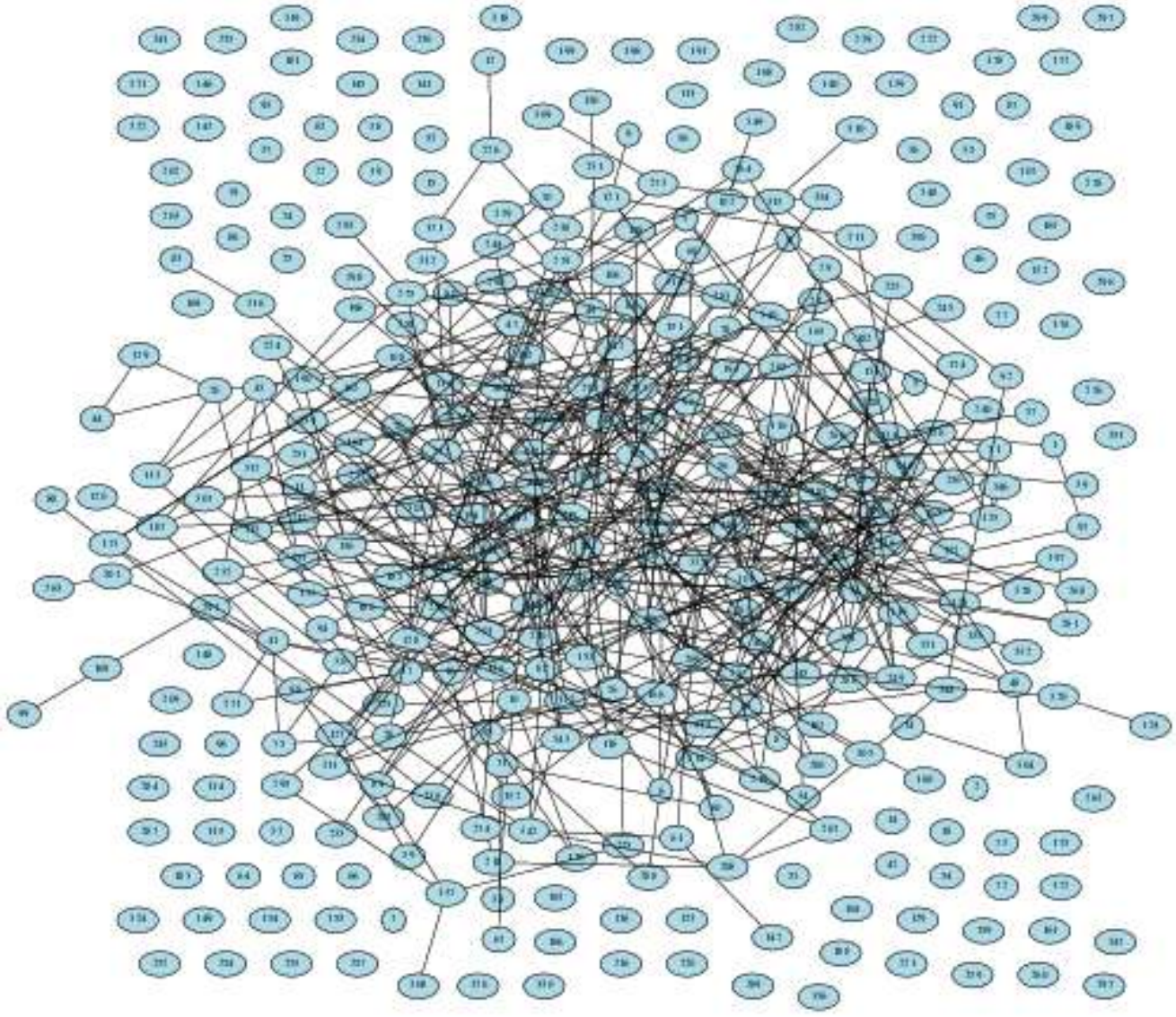


Figure 8.9 Non-Featured Article: *Timber Framing* Editing Network

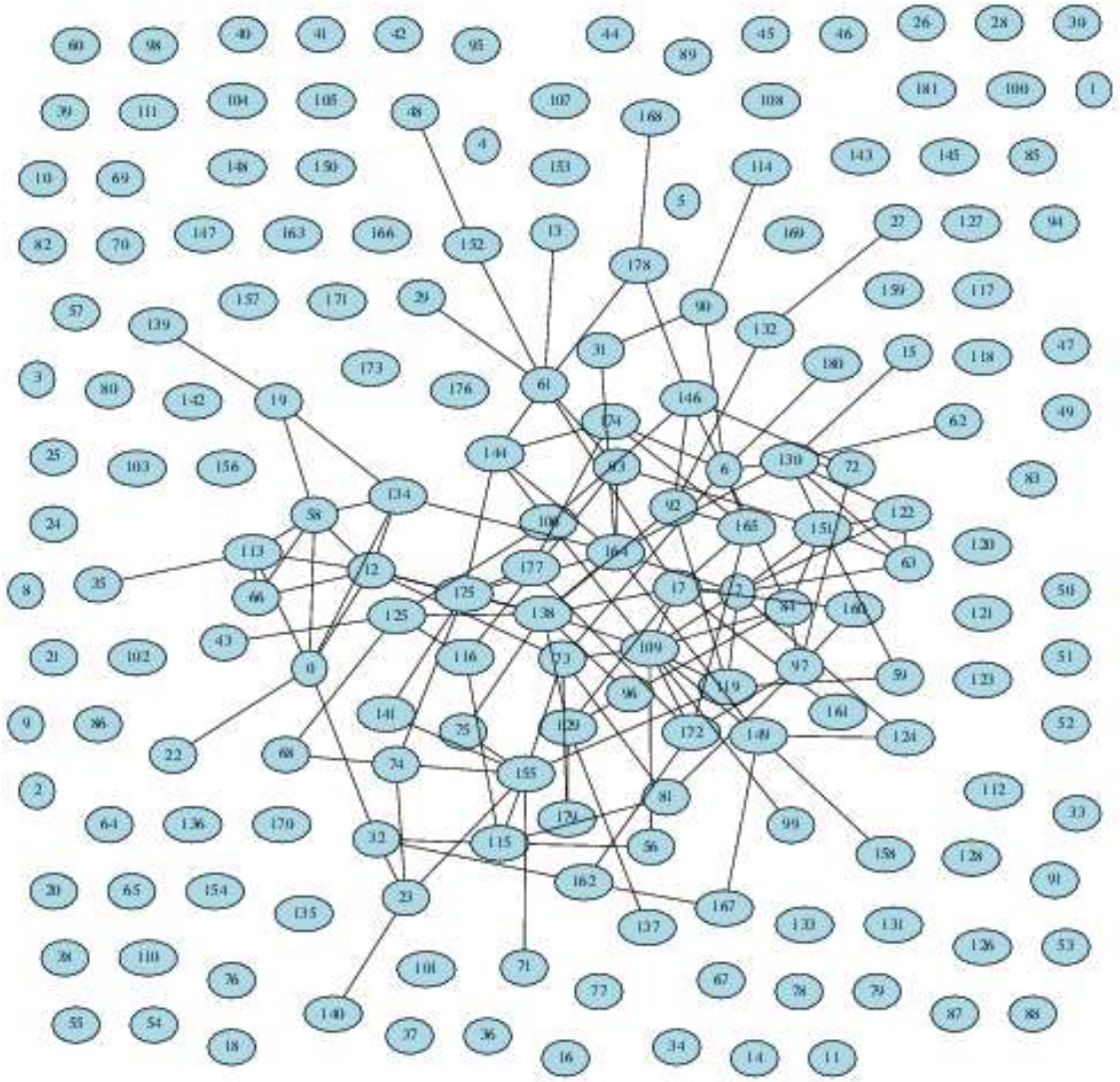


Figure 8.10 Non-Featured Article: *Flooding of the Nile* Editing Network

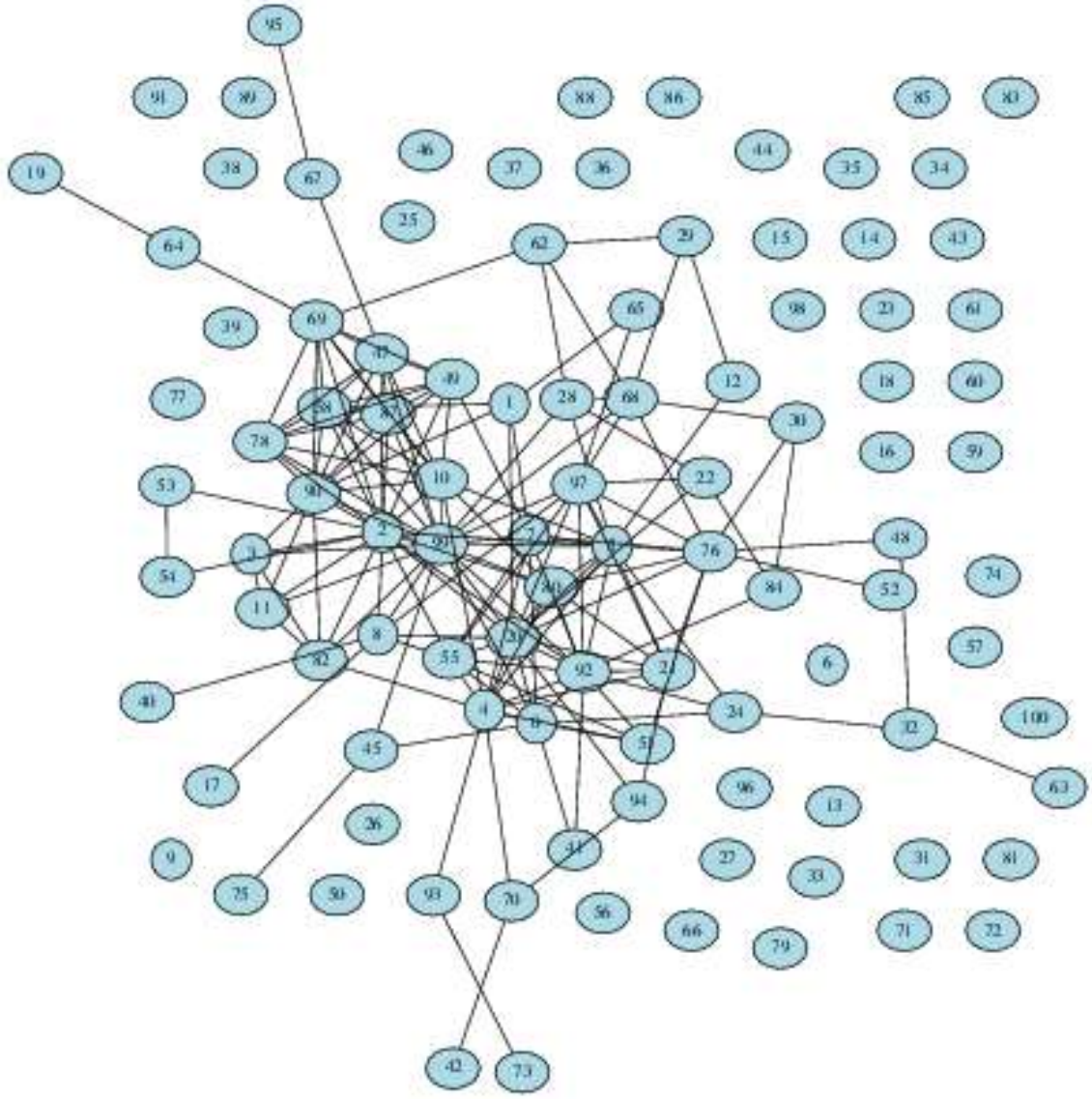


Figure 8.11. Non-Featured Article: *Stabilizing Selection* Editing Network

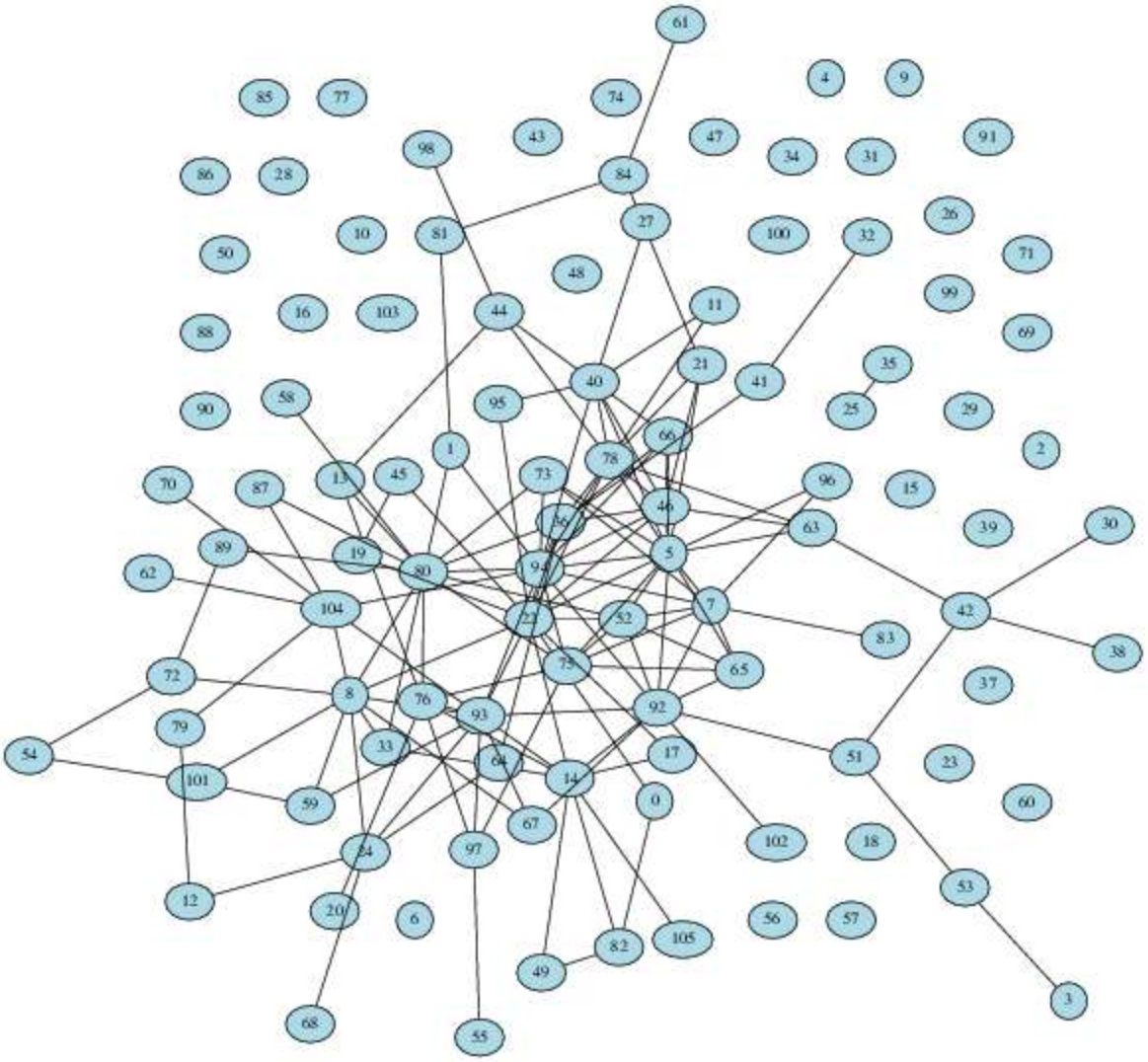


Figure 8.12. Non-Featured Article: *Systematics* Editing Network

References

[1] Welser, H., Kossinets, G., Marc, S., and Cosley, D. Finding Social Roles in Wikipedia. *Paper presented at the annual meeting of the American Sociological Association*, Boston, MA. (2008).

[2] Mccallum, A., & Wang, X. Topic and Role Discovery in Social Networks. *In Proceedings of 19th International Joint Conference on Artificial Intelligence*, 2005.

[3] Leskovec, J. (2010). Kronecker Graphs : An Approach to Modeling Networks. *Journal of Machine Learning Research, 11*, 985-1042.

[4] Shi, X., Leskovec, J., & McFarland, D. a. (2010). Citing for high impact. *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*, 49. New York, New York, USA: ACM Press.

[5] Lih, A. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *In Proceedings of the 5th International Symposium on Online Journalism*, (April, 2004), Austin, USA

[6] Stvilia, B., Twidale, M.B., Smith, L.C. and Gasser, L. Assessing information quality of a community-based encyclopedia. *In Proceedings of the International Conference on Information Quality*, 442–454, (November, 2005), Cambridge, USA

[7] Adler, B.T. and de Alfaro, L. A Content-Driven Reputation System for the Wikipedia. *In Proceedings of the 16th International Conference on the World Wide Web*. 261- 270, (May, 2007), Banff, Canada.

[8] Wöhner, T., & Peters, R. Assessing the quality of Wikipedia articles with lifecycle based metrics. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration - WikiSym '09*, 1. New York, New York, USA: ACM Press.

[9] Blumenstock, J.E. Size Matters: Word Count as a Measure of Quality on Wikipedia. *In Proceedings of the 17th international conference on World Wide Web.* 1095-1096, (April, 2008). Beijing, China.

[11] Leskovec, J. Dynamics of large networks, Ph.D. Dissertation, Machine Learning Department, School of Computer Science, Carnegie Mellon University, Technical report CMU-ML-08-111, September

[12] Stanford Network Analysis Platform. http://snap.stanford.edu/snap/

[13] A. –L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[14]. S. L. Tauro, C. Palmer, G. Siganos, and M. Faloutsos. A simple conceptual model for the internet topology. In *GLOBECOM '01: Global Telecommunications Conference*, volume 3, pages 1667– 1671, 2001.

[15] A. Clauset, C. R. Shalizi, andM. E. J. Newman. Power-law distributions in empirical data. ArXiv, ArXiv:0706.1062, Jun 2007

[16] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In KDD '05: Proceeding of the 11th ACM SIGKDD *International Conference on Knowledge Discovery in Data Mining*, pages 177–187, 2005.

[17] Watts, Duncan J.; Strogatz, Steven H. (June 1998). Collective dynamics of 'small-world' networks. *Nature* **393** (6684): 440–442.

[18] A. Clauset, C. R. Shalizi, andM. E. J. Newman. Power-law distributions in empirical data. ArXiv, ArXiv:0706.1062, Jun 2007.

[19] Iba, T., Nemoto, K., Peters, B., & Gloor, P. a. (2010). Analyzing the Creative Editing Behavior of Wikipedia EditorsThrough Dynamic Social Network Analysis. *Procedia - Social and Behavioral Sciences*, *2*(4), 6441-6456

[20] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[21] Kim, M., And Leskovec, J. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. *SDM* (2011).

[22] Tong, H., Prakash, B. A., Tsourakakis, C., Rad, T. E., Faloustos, C., And Chau, D. H. On the Vulnerability of Large Graphs. In *ICDM* '10 (Los Alamitos, CA,USA, 2010), vol. 0, IEEE Computer Society, pp. 1091–1096.

[23] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. *In ICML '07: Proceedings of the 24th International Conference on Machine Learning*, 2007.