Bipartite Networks of Wikipedia's Articles and Authors: a Meso-level Approach

Rut Jesus
Center for Philosophy of Nature and Science Studies
University of Copenhagen
Blegdamsvej 17, 2100 Copenhagen, Denmark
+4561339903
vulpeto@gmail.com

Martin Schwartz
IT University of Copenhagen, DK-2300 Copenhagen S and Informatics and Mathematical Modelling. Technical University of Denmark. DK-2800 Kgs. Lyngby, Denmark
+45 50571799
the1schwartz@gmail.com

Sune Lehmann
Center for Complex Network Research and Department of Physics, Northeastern University, Boston and Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Harvard University, Boston, MA 02115, USA
+1(617)3738806
sune.lehmann@gmail.com

ABSTRACT

This exploratory study investigates the bipartite network of articles linked by common editors in Wikipedia, 'The Free Encyclopedia that Anyone Can Edit'. We use the articles in the categories (to depth three) of Physics and Philosophy and extract and focus on significant editors (at least 7 or 10 edits per each article). We construct a bipartite network, and from it, overlapping cliques of densely connected articles and editors. We cluster these densely connected cliques into larger modules to study examples of larger groups that display how volunteer editors flock around articles driven by interest, real-world controversies, or the result of coordination in WikiProjects. Our results confirm that topics aggregate editors; and show that highly coordinated efforts result in dense clusters.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Computer-supported cooperative work, Web-based interaction*; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*; J.4 [**Social and Behavioral Sciences**]: Miscellaneous.

General Terms

Algorithms, Design, Human Factors.

Keywords

Bicliques, Wikipedia, Collaboration, Meso-level.

WikiSym '09, October 25-27, 2009, Orlando, Florida, U.S.A.

1. INTRODUCTION

Wikipedia is a good example of social production of knowledge. Authors and articles constitute a network, which we study here at the meso-level. Investigations on knowledge-producing agents and their networks are of interest to both network and quantitative analysis studies, as well as to the social sciences. Moreover, it is particularly interesting to try to understand the network structure and dynamics inferred from low level information subsequentlyl complemented with higher level information. Wikipedia's network of authors and articles, is more horizontal than other networks (for example, those of the peer-reviewed scientific literature) – e.g., it has more edits per person and per article.

1.1 Related Literature

1.1.1 Network analysis

Network analysis has previously been used to describe Wikipedia's growth. For instance, Capocci *et al.* (2006) [1], delineate the properties of the growth of Wikipedia as a network, with topics modeled as vertices and hyperlinks between them represented as edges. This study shows how the growth of Wikipedia can be described with local rules such as preferential attachment, while contributors are still free to act globally in the network. It has also been discovered that many network characteristics are similar between different language versions of Wikipedia; examples are degree distribution, growth, reciprocity and clustering, Buriol *et al*, 2006[2]; Zlatic *et al*, 2006 [3]).

1.1.2 Quantitative analysis

Quantitative analysis of Wikipedia users has been investigated by Ortega and Gonzalez-Barahona (2007) [4] in a framework, where editors were classified by their activity during specific time periods. A comparison between imposed classifications and real clustering was performed by Capocci, Rao and Caldarelli (2008) [5].

1.1.3 Cooperation

The level of cooperation in Wikipedia has been carefully analyzed by Viegas et al. (2007) [6], who stress the need to study Wikipedia's growth in terms of its clusters and namespaces beyond the articles. These authors emphasized that the fastest growing areas (namespaces) in Wikipedia are devoted to coordination of article-writing and conventions. They create a grid of categories and code the contents of discussion pages according to that grid. They discover that these pages mostly act as a place for strategic planning of edits and enforcement of standard guidelines. A study by Wilkinson and Huberman (2007) [7] shed light on the stochastic mechanism by which articles accrete edits. They show that there is a positive correlation between article quality and number of edits, thereby validating Wikipedia as a successful collaborative effort. A more recent study by Kittur and Kraut (2008) [8] specifies better the impact of adding editors for the quality of articles: the addition of editors improves the quality of an article in its formative stage, and when the coordination is done directly in the writing of the article, but the addition of editors to an article can be harmful when the coordination is done explicitly in talk pages.

1.1.4 Visualizations

The visualization of collaboration within Wikipedia is also an active field of research; the tools include: (1) *history flow* (Viégas *et al*, 2004 [9]) an application that can be used to visualize the contributions to an article; (2) visualization of the whole co-authorship networks (Biuk-Aghai, 2006 [10]), and (3) the use of revert graph visualizations (Suh *et al*, 2007 [11]).

1.2 Conceptual Framework

1.2.1 Meso-zoom

Most of the work referenced above focuses on either the global statistics of the entire Wikipedia project, or on the atomic descriptions of individual articles. However, collaboration in Wikipedia occurs at the meso-level, where groups of people collaborate in order to create articles . We here focus on the meso-level, not only in terms of scale, but also in terms of analysis. This is a study where low-level phenomena – *i.e.,* agents and their interactions and behaviors, inform a higher level – that of clusters between articles and editors.

1.2.2 Meso-approach

We stay in the middle. Modules of articles and editors are investigated, rather than whole wikipedias and their statistics or single discussions and their descriptive sociologies. Moreover, we stay in the middle regarding our approach, supported by our inter-disciplinary skills in physics and philosophy: we employ network visualizations, but we neither make comprehensive statistical analyses nor detailed ethnographic studies. Although this interdisciplinary approach may appear lacking from the point of view of either of these 'pure fields', we believe that the inter-disciplinary nature of this study allows us to integrate mathematical tools and sociological methodologies to allow us to see general patterns without the oversimplification that is often the result of a purely quantitative approach.

In the following sections we introduce bipartite networks and present and defend our choices concerning data and visualization. Subsequently we show several case studies and examples of bipartite modules surrounding various controversies, interests and projects. We consider the network formed by overlapping clusters of articles and editors and utilize this to detect isolated cliques. We also present the clusters, which are not bounded by content. Finally, we discuss the results and propose lines for future research.

2. Method

2.1 Bipartite Networks

A bipartite network is a graph $G = (U, V, E)$ whose vertices (or 'nodes') can be divided into two disjoint sets U and V such that every edge (or 'link') E connects a vertex in U to a vertex in V; that is, U and V are independent sets. When we consider articles in Wikipedia and their editors, a bipartite network is a convenient representation: U is the set of editors and V is the set of articles in Wikipedia. The bipartite network formalism is ideal for studying collaboration, because the network structure encodes knowledge about which articles editors have edited together.

By studying the clusters (or 'modules') in the bipartite network, we are able to discover clustering of editors and articles and smaller patterns of collaboration. We choose to call dense groups clusters or modules rather than 'community', because the latter is an ill-defined concept across disciplines and may imply structures at the macro-level not present in this meso-level study. These dense groups could also be called 'epistemic communities' as used by Roth (2006) [12] where epistemic communities are understood as a descriptive instance only, not as a coalition of people who have some interest to stay in the community: it is a set of agents who participate in building the same knowledge.

Bicliques or their various names (closed sets, closed couples, formal concepts, maximal rectangles, bipartite communities) were initially studied by mathematicians Birkhoff (US), Barbut (F) together with Monjardet (F) and by computer scientist Rudolf Wille (DE). And they continue to be explored in formal concept analysis and by mathematical sociologists. We do not choose to review these mathematical formulations of bicliques at length, and focus instead on their use in network research where they are the building block of clusters/communities/groups.

One method for detecting modules in bipartite networks, grounded in physics of networks and expanding the work by Palla *et al* (2005) [13] was developed by Lehmann *et al.* (2007) [14]. This method is based on detecting the most dense areas of the graph (called maximal bi-cliques) and then agglomerating overlapping bi-cliques into larger modules. More formally, a biclique is a complete subgraph of a bipartite network. A 'maximal' biclique is defined as a biclique that is not a subgraph of any larger bi-clique. We use the notation $K_{u,v}$ to describe a bi-clique with u nodes in node-set U and v nodes in node-set V.

Connecting this to the network of editors and articles in Wikipedia, a $K_{3,5}$ cliques describes a structure where three editors have all edited the same five articles. Two bi-cliques of size $K_{a,b}$ are adjacent if they share at least a $K_{a-1,b-1}$ clique. A $K_{a,b}$ module (or 'cluster') is the union of all adjacent $K_{a,b}$ cliques. One important feature of this definition is that nodes can belong to more than one cluster; that is, two distinct modules may overlap. Furthermore, by changing the values of a and b allows for different zooms.

2.2 Data and Visualization

2.2.1 Subset

We analyze a subset of the English language Wikipedia, namely the articles in the categories Philosophy and Physics to depth level three. The choice of a subset is, after all, arbitrary but our sample was motivated by familiarity with the topics (given our educational background, which is important to make semantic claims about them) and by the size of the disciplines and their representation in Wikipedia. As categories in Wikipedia can be nested recursively, the set of articles includes not only articles inside Physics and Philosophy but also those in different subjects up to three steps of association from the main categories. The decision to include sub-categories and sub-sub-categories is similar to the choice of Halavais and Lackaff (2008) [15]. The authors assume that a 'core' of the disciplines can be sampled in this way.

2.2.2 Filtering

Similarly, editors were filtered by the number of edits they had contributed to each article. Editors that edited 7 or more, or 10 or more times in an article were included (both thresholds were applied but for different purposes). This filtering helped to avoid clutter (and allow for computational capacity), and helps us concentrate on the most engaged editors and articles in dense clusters. Although sporadic edits can be important to Wikipedia as a whole, they are less relevant when considering the cooperation and interaction between editors of a small subset of articles. A few examples indicate that the lack of information regarding sporadic editors does not compromise the analysis of highly engaged clusters of editors and articles.

2.2.3 Anonymity

The 'real nicknames' of the editors are kept due to the public nature of their work; as is clear from the examples, their identity is not at stake, not more than by creating an account in the Wikipedia website.

2.2.4 Bi-clique visualization

The open source program, BCFinder developed by Lehmann *et al* (2007) [14] was used to calculate and visualize the modules that arise from combining adjacent bi-cliques. BCFinder allows one to visualize the articles and editors of each cluster (and also easily access those pages and user pages in Wikipedia)., In addition, it makes it possible to visualize the network of modules. Each $K_{a,b}$ module can be thought of as 'zooming' into a relevant area of the network. Moreover, one can view the network of modules, where each cluster is a node and two module-nodes are linked if they share either one or more editors or one or more articles (see example in Fig. 5). A distinct network of modules is created by each zoom and yields insight about which zooms divide the network into meaningful sub-parts. Further, the network of clusters allows one to identify isolated clusters. Using the filter of minimum number of edits-per-article-per-editor set to 7, we worked with 33335 editors and 17643 articles (we call this the '7-edit network'); when the minimum number of edits-per-article-per-editor was set to 10 we worked with 19612 editors and 13241 articles (the '10-edit network').

2.2.5 Typology

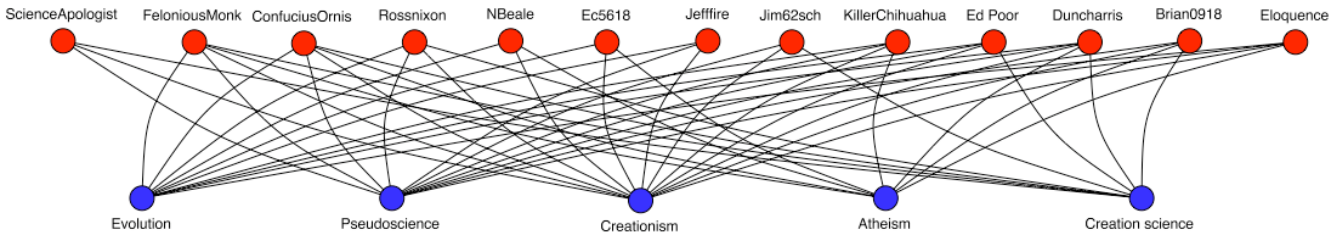
Upon getting all possible clusters, they were grouped in order to identify the specific examples below. Although taken from a specific $K_{a,b}$ cluster, these examples are fairly robust to changes in *a* or *b,* up to a certain point. The kinds shown below span the possible types found in the data.

3. RESULTS

3.1 Controversies

3.1.1 Evolution/Creationism

The first type of collaboration in Wikipedia is the one fueled by deep disagreement. One example of such a cluster is the controversy between evolution and creationism. In Figure 1 we display the major players of this cluster tying controversial articles. Here, we study the 10-edit network. The module is composed of adjacent bi-cliques of size $K_{6,3}$ or greater. The articles present in this cluster show that a debate is taking place. For example, the two articles 'Evolution' and 'Creationism', are edited by the same group of editors. The controversy here surrounds a religious/non-religious discussion that ultimately questions the validity of science. The presence of 'Atheism' and of 'Pseudoscience' supports the debate of religious values in relation to scientific values.



.

Figure 1: The Evolution/Creationism debate is mirrored in the way the articles 'Evolution', 'Pseudoscience', 'Creationism', 'Atheism' and 'Creation science' belong to the same cluster. These articles are edited by at least 13 active editors engaged in this controversy.

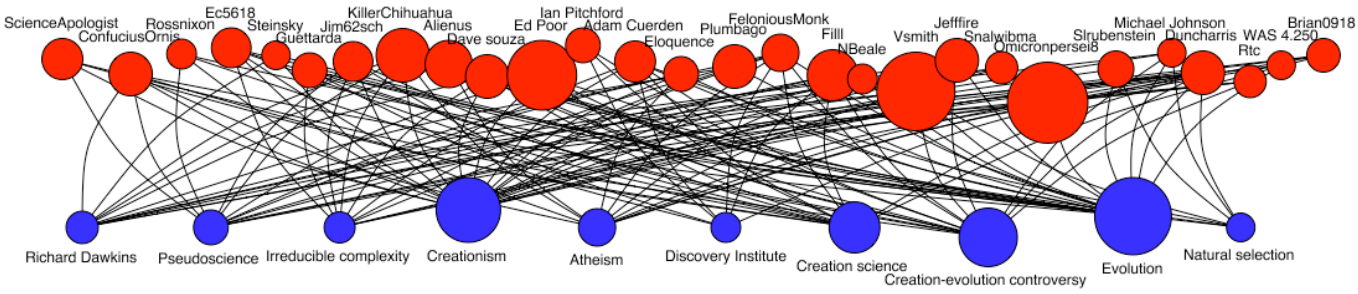


Figure 2: Zooming in the Evolution/Creationism debate by including more edits. The vertices are scaled according to number of links. More articles and more editors are involved in this dispute. This cluster gives clues about some of the hidden players, for example 'Richard Dawkins' and the 'Discovery Institute'.

In Figure 2, another cluster around the same topic is shown. In this figure, each vertex is scaled such that nodes with more links are larger; this makes it easier to see that the two major articles are 'evolution' and 'creationism'. Some of the smaller articles yield further insight into other actors participating in this dispute: 'Richard Dawkins' "is a British ethologist, evolutionary biologist and popular science writer. In addition to his biological work, Dawkins is well-known for his views on atheism, evolution, creationism, intelligent design, and religion. He is a prominent critic of creationism and intelligent design" as is stated in the first lines of the Wikipedia article[i]. An important concept in this controversy seems to have been heavily edited as well: 'Irreducible complexity' which "is an argument made by proponents of intelligent design that certain biological systems are too complex to have evolved from simpler, or "less complete" predecessors, through natural selection acting upon a series of advantageous naturally occurring chance mutations"[ii]. On the other side of the debate, the major concept at stake is 'Natural Selection' which "is the process by which favorable heritable traits become more common in successive generations of a population of reproducing organisms, and unfavorable heritable traits become less common."[iii] These clusters can also reveal players that would be otherwise hidden to those not involved. For example, the Discovery Institute "is a U.S. think tank based in Seattle, Washington, best known for its advocacy of intelligent design and its Teach the Controversy campaign to teach creationist anti-evolution beliefs in United States public high school science courses."[iv] Investigating the other set of nodes (editors) involved in this discussion is also revealing. Represented in their user pages we discover a range of attitudes. One editor states clearly that he was involved with the article 'Intelligent Design', which he started in 2001, but from which he was banned in 2008. Other editors decided to leave Wikipedia—it is not clear if the controversy discussed here played a role. Still other editors appear to have been highly involved in fighting vandalism; it is well known that controversies are more prone to vandalism (Viégas *et al*, 2004 [9]).

3.1.2 Intelligence and Global Warming

Several other controversies can be identified based on the modules in our subsection of Wikipedia. The controversy in Figure 3 is based on the 7-edit network, and displays a module based on $K_{5,4}$ bi-cliques. This group is engaged in a discussion of the issue of intelligence and the validity of the intelligence tests and some claims for correlations. In addition, 'The Bell Curve' is a controversial book on how intelligence can be a predictor of social factors. Likewise 'IQ and the Wealth of Nations' is another controversial book discussing the relation between IQ prosperity of nations.

[ii] From Wikipedia, "Richard dawkins." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Richard Dawkins.

[ii] From Wikipedia, "Irreducible complexity." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Irreducible complexity.

[iii] From Wikipedia, "Natural selection." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Natural selection.

[iv] From Wikipedia, "Discovery institute." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Discovery institute.

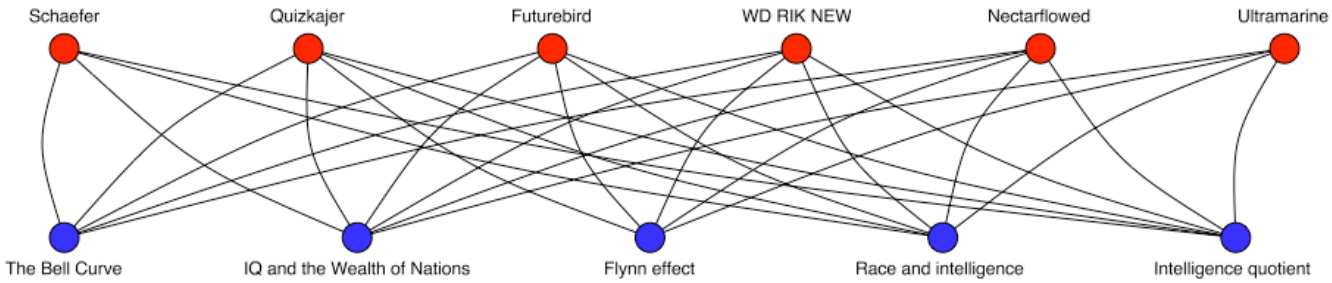


Figure 3: Controversy surrounding intelligence, its measures and correlations comprised of the articles 'Race and intelligence', 'The Bell Curve', 'IQ and the Wealth of Nations', 'Intelligence quotient', and 'Flynn effect'.

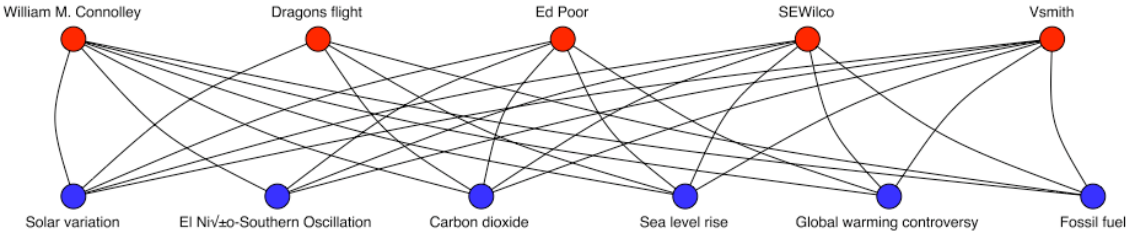


Figure 4: Controversy surrounding global warming. It comprises the articles 'Solar variation', 'El Niño-Southern Oscillation', 'Carbon dioxide', 'Sea level rise', 'Global warming controversy', and 'Fossil fuel'.

Another characteristic example of a 'conflict-cluster' is displayed in Figure 4. Here controversy regards global warming and the diverse factors surrounding this subject. The network is based on the 7-edit filter and the module is slightly more sparse than the ones considered so far, constructed from adjacent $K_{4,3}$ bi-cliques. The central article in this cluster is 'Global warming controversy'. But the pages 'Solar variation', 'Carbon Dioxide', 'Sea level rise', 'Fossil Fuel' and 'El Niño-Southern Oscillation' are all components in the discussion on the human components involved in global warming.

3.2 Isolated Clusters

In order to understand the significance of the next type of collaboration in Wikipedia, it is useful to first discuss the network of modules. The network of modules allows one to identify modules in the bipartite network of editors and articles, which are not connected to any other modules. Figure 5 is an example of the network between the modules 10-edit network, with modules constructed from $K_{7,2}$ cliques. Each module is represented by a pie-chart colored according to its fraction of editors (red) and articles (blue). The modules are connected by red links (overlapping editors) and blue links (overlapping articles); the width of each link is proportional to the number of overlapping nodes.

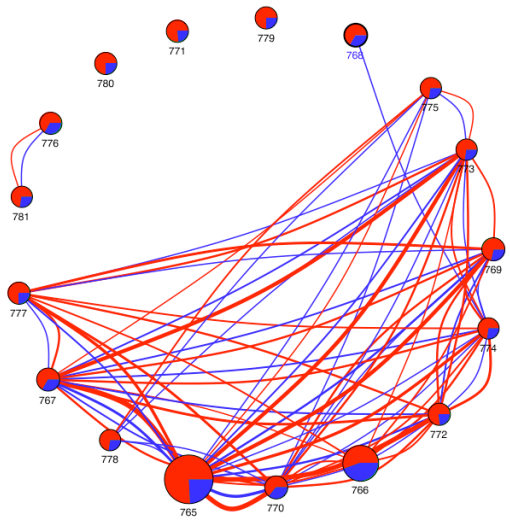


Figure 5: Network of the clusters made of K $_{7,2}$ bi-cliques. Circles represent modules, which share articles (blue links) and editors (red links) with each other. The numbers are labels that identify each cluster. The network-of-clusters-view helps to understand the relationships between the clusters and to identify isolated clusters that do not share articles or editors with others. The clusters 780, 771, and 779 are displayed in Fig. 6; the clusters labeled 776 and 781 are displayed in Fig. 7.

Figure 5 shows three clusters 780, 771 and 779 (these numbers are just labels) that do not share links (either articles or editors) with the others. Two other clusters 780 and 776 are sparsely

connected. Let us investigate these modules and begin to understand the causes underlying this network topology. The three isolated clusters correspond to topics that gather focused and dedicated authors: Mormonism, Zionism and Scientology (Figure 6). It is not fully surprising that all of those topics are isolated from other clusters since it could be argue that their practice in the 'real world' is similar: organized in sub-cultures, highly active, but isolated from other areas of knowledge and/or society. In Figure 5, two other clusters are connected with each other but not with the remaining modules; these are plotted in Figure 7. Both modules are devoted to political 'isms' and share one editor and a single article the one on 'Anarchy'. One of these clusters is interested in the definition and background of anarchism as the articles are: 'Individualist anarchism', 'Mutualism (economic theory)' (is an anarchist school of thought) and 'Anarchism'. This cluster is then related to another interested in defining political 'isms': 'Anarchism', 'Anarcho-capitalism', 'Socialism' and 'Capitalism'.

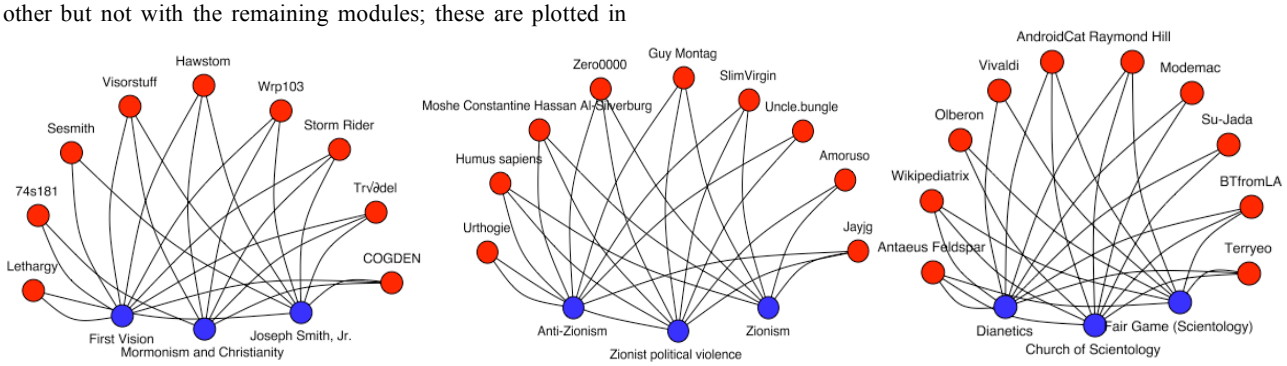


Figure 6: Isolated clusters: The left panel is a module focused on the topic of Mormonism, which comprises paradigmatic articles: 'First Vision', 'Mormonism and Christianity' and 'Joseph Smith, Jr.'; the middle panel surrounds the topic of Zionism in all three articles: 'Anti-Zionism', 'Zionist political violence' and 'Zionism'; the right panel surrounds the topic of Scientology: 'Dianetics', 'Church of Scientology' and 'Fair Game (Scientology)'.

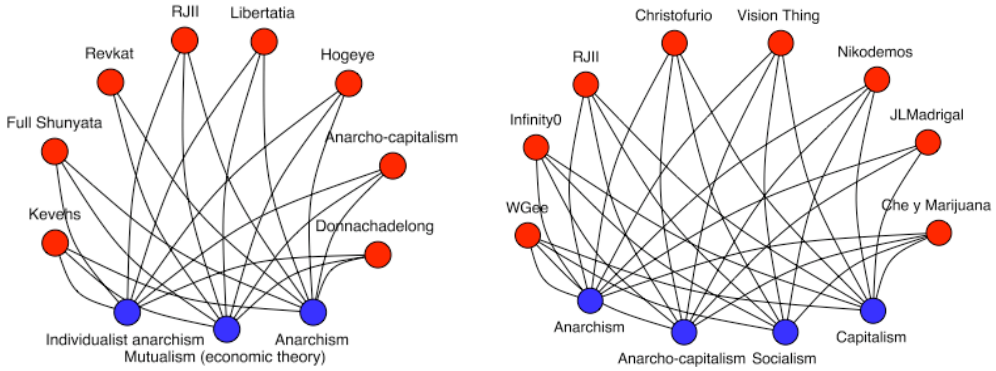


Figure 7: Two connected clusters that are disconnected from the remaining network of modules. (left) Cluster focused on Anarchism. (right) Cluster focused on political 'isms': 'Anarchism', 'Anarcho-capitalism', 'Socialism' and 'Capitalism'.

3.3 Shared Interests

In all the clusters, the editors share the interest (and practice) of editing the same articles. Some of them can be grouped by a shared interest (or a number of related ones). These groups are revealed by the bi-cliques, some of which turn out to be coordinated through a WikiProject.

3.3.1 Mantras

Figure 8 shows another example of a cluster realized from shared interest practice, although this one is not concentrated in a WikiProject. This project concerns the topics 'Buddhism', 'Yoga', 'Tantra', 'Mantra' and 'Guru'. It reveals common interests between the 5 editors and the 5 articles in this $K_{4,4}$ bi-clique cluster based on the 7-edit network. Although 'Tantra', 'Yoga' and 'Guru' are not related directly, they are part of the same vocabulary and interests of the practitioners of yoga, guru followers and tantra interested people. This cluster reflects a practice that happens beyond Wikipedia, but a practice that is mapped onto the way the articles are edited.

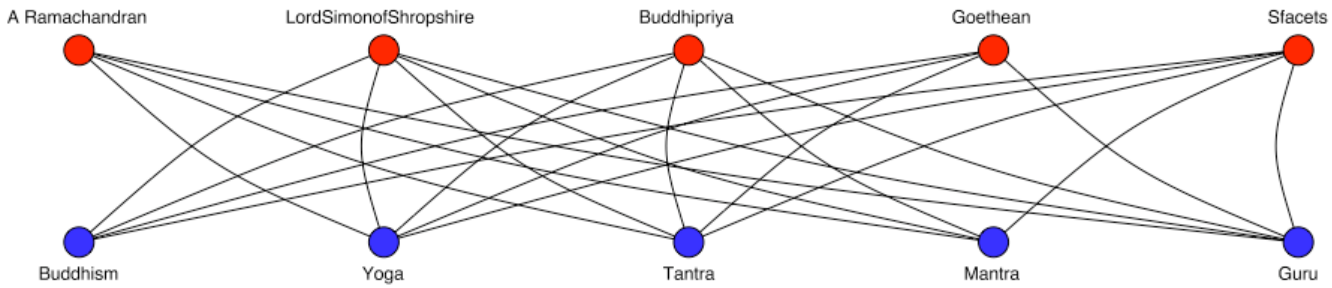


Figure 8: Cluster showing a relation between articles about related practices: 'Buddhism', 'Yoga', 'Tantra', 'Mantra', and 'Guru'.

3.3.2 WikiProjects

3.3.2.1 Elements

Figure 9 displays 8 articles and 10 editors, which constitute a $K_{7,3}$ module in the 10-edit network. This collaboration is a clear example of an orchestrated effort to improve the articles describing the elements of the periodic table. One of the WikiProjects is "a collection of pages devoted to the management of a specific topic or family of topics within Wikipedia; and, simultaneously, a group of editors that use said pages to collaborate on encyclopedic work. It is not a place to write encyclopedia articles directly, but a resource to help coordinate and organize article writing and editing"[v]. The WikiProject about elements presents itself in the following manner: "This WikiProject has managed to standardize the articles on the known chemical elements (see Guidelines page). Now it is aimed at the maintenance of these at an agreed upon format discussed in Wikipedia talk:WikiProject Elements and at the expansion and improvement of each article to featured article quality (check out our Goals below)."[vi]. In this cluster the editors are engaged in improving the following articles: 'Hydrogen', 'Oxygen', 'Gold', 'Mercury', 'Magnesium', 'Lithium', 'Krypton', 'Potassium'. An investigation of their user pages reveals that the editors involved are several administrators with daily activities that range from working mathematicians to geologists and chemists. The various editors have different levels of (dis)comfort with anonymity: some use their real name, some keep it hidden but provide extensive information about their activities and, at least one, copes with anonymity in an interesting manner: "Male, European, and already paranoid about giving away this much information".

[v] From Wikipedia, "Wiki project." Retrieved on May 2nd, 2008 from http://en.wikipedia.org/ wiki/Wikipedia:WikiProject.

[vi] From Wikipedia, "Wikiproject elements." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Wikipedia:WikiProjectElements.

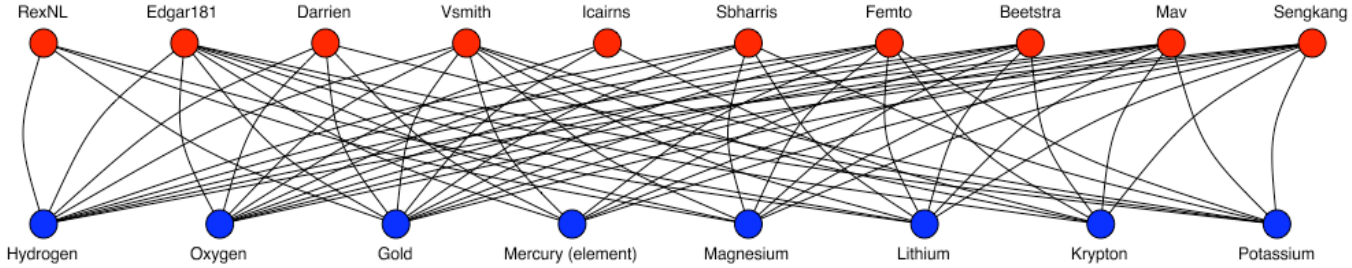


Figure 9: Cluster revealing the coordinated effort to improve Wikipedia articles about the elements of the Periodic Table.

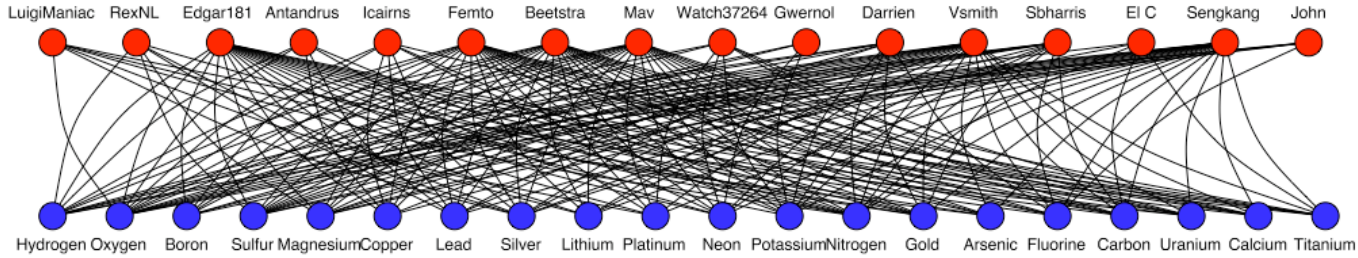


Figure 10: Cluster showing more elements that are part of the WikiProject concerned with improving the articles of the elements of the Periodic Table by decreasing the minimum number of edits allowed.

Additional data about this cluster can be obtained by considering the 7-edit network. For the same clique zoom of $K_{7,3}$, Figure 10 has 16 editors and 20 articles. As it is a coordinated effort, the additional information gained by increasing the number of edits is only that there are more people and articles involved in the same topic: We see 20 elements instead of the 8 elements that

were visible in the case of the previous cluster with fewer editors and articles.

3.3.2.2 Electronics

Another example of a cluster that reveals a WikiProject is displayed in Figure 11. This project surrounds the topic of electronics and several of its concepts ('Alternating current', 'Decibel') and tools ('Oscilloscope', 'Electric motor'). The $K_{4,4}$ clique cluster comprises 7 editors and 11 articles in the 7-edit network. The presentation of the WikiProject about Electronics is the following: "The aim of this project is to better organize information in articles related to electronics. This page contains only suggestions, with the hope to help other Wikipedians writing high-quality articles with the minimum effort"[vii].

[vii] From Wikipedia, "Wikiproject electronics." Retrieved on May 2nd 2008 from http://en.wikipedia.org/wiki/Wikipedia:WikiProjectElectronics.

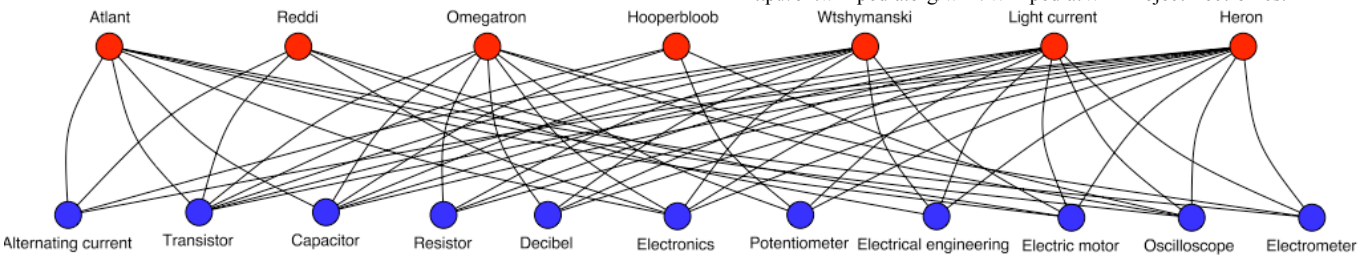


Figure 11: Cluster supported by the WikiProject Electronics around the topic of, well, of electronics: 'Electrometer', 'Decibel', 'Potentiometer', 'Alternating current', 'Electrical engineering', 'Electronics', 'Oscilloscope', 'Resistor', 'Transistor', 'Electric motor' and 'Capacitor'.

3.4 Non-Content Bounded Clusters

The bipartite network of editors and articles also contains modules, in which there is no apparent correlation between the topics. Figure 12 displays such a module with 5 articles and 9 editors around topics as diverse as: 'Joseph Stalin', 'Martin Luther King, Jr.', 'Tsunami', 'Ku Klux Klan' and 'Albert Einstein'. It is a curiosity to observe what topics would be included in these generalist clusters that are heavily edited and by a small group of editors.

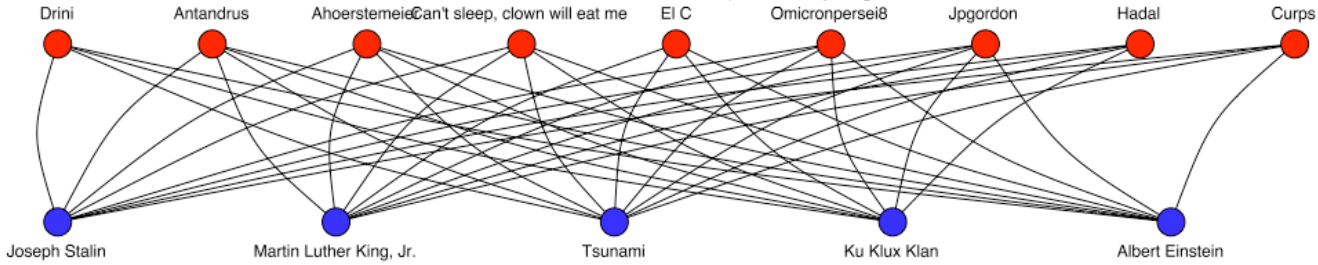


Figure 12: There are also several clusters such as this one, which are not bounded by content, but probably by editing style edits - maybe for adding links or fighting vandals.

A more extensive list from a module of adjacent $K_{4,1}$ bi-cliques with 45 articles is: Abortion, Jimmy Wales, Solar energy, Evolution, Fuck, Christianity, Ku Klux Klan, Beauty, Galileo Galilei, Racism, Stupidity, Black hole, Plato, Joseph Stalin, Sun, Volcano, Aristotle, Earthquake, Art, Rosa Parks, Nuclear power, Isaac Newton, Computer, Martin Luther King, Jr., Tsunami, Buddhism, Creationism, Bitch, Vietnam War, Tornado, Pi, Shit, Pope John Paul II, Albert Einstein, Internet, Thomas Jefferson, Vladimir Lenin, Love, Cunt, Renaissance, Islam, Slavery, Mother Teresa, Tropical cyclone, Music. One possible way to account for this variety in topic in this example of a cluster not bounded by content is that these articles have very general content, they are not highly specialized and therefore are more accessible to different kinds of editors. Another complementary explanation is that articles are sometimes edited, not in terms of topic, but rather kind of edit. An editor that is concerned with making tables, or fixing links would not be concerned with the specific topic and the edits are therefore due to syntax, layout, or spelling editors.

4. DISCUSSION

By applying clustering tools from social network analysis to a subsection of Wikipedia, several interesting insights regarding the meso-level between single articles and global statistics were uncovered. Although we were limited to the articles that were included in the subsections of the categories Physics and Philosophy and therefore related to these two primary topics, these boundaries gave us a certain familiarity with the topics. This facilitated the extraction of information in a manner that would not have been possible, had the research been performed on random or unfamiliar topics.

Controversies give rise to disputes that are not necessarily contained within one article. In fact, controversies typically span multiple articles and form tightly connected modules of editors who edit related topics actively and sometimes in direct opposition to each other. Wikipedia, as expected, mirrors the discussions in society. Clustering tools allow us to probe other structures than the 'web of knowledge' that arises from the networks where the nodes are articles and the hyperlinks connect them.

The article on 'Evolution' links not only to 'Darwin' or 'Wallace', but also connects to 'Atheism', for example. This modular structure reflects the controversy currently taking place on the scale of the entire North-American society, which is actively engaged in discussing the possibility of creationism to be taught alongside with evolution. In this manner, analyzing the modules in Wikipedia, provides information about another layer of the construction of knowledge which is not necessarily tied with the topics closest in character, but with those that create issues which must be articulated and disputed in relation to each other.

In the case of the coordinated efforts, such as the Project Elements, the attempt to achieve Featured Article status seems to aggregate people (Viegas, Wattenberg and McKeon, 2007) [16] and also, as previously proven by Wilkinson and Huberman (2007) [7] the more edits an article has, the more it is likely to accrete. Therefore, the creation of WikiProjects is shown to be a good way to mobilize work in one direction, especially by trying to produce Featured Articles. WikiProjects are a good example of the work carried out at the meso-level: they do not rely on massive inputs by the 'wisdom of the crowds' nor do they rely uniquely on the dedication of one single editor. WikiProjects result in clusters of editors with common interests that have found a way to coordinate work successfully aggregating people and resulting in highly developed articles. As expected, some clusters reflect the way those same clusters manifest in 'real life'. If topics or practices aggregate tight and closed clusters, it is not surprising that the articles about those clusters are also edited by a closed cluster of editors.

The bipartite clustering tools and the network of modules can be used, not only to identify some of those modules in 'real life', but also to understand the relations between the modules and the most important players. For example, in the controversy between 'Evolution' and 'Creationism', there are people and groups who are quite outspoken ('Dawkins', 'Discovery Institute') and therefore their articles are edited along with the other articles present in the controversy. Another example is that specific properties about how some articles are edited can be related to some assumed properties of groups in the 'real world': the isolated groups on 'Mormonism', 'Scientology' and 'Zionism' may show that these groups in society are also quite isolated and dedicated to their cause. Although it is hard to prove the behavior of these groups in 'real life', a recent case of Wikipedia banning the Church of Scientology from editing (Wired, 2009) [17] supports that these editing patterns may reflect that these groups edit directly their own pages and that the discussions about them, some even controversial, are quite isolated.

These clusters can be seen as 'epistemic communities', in the weak sense, that of a group of people gathering around a knowledge topic (and not in the strong sense where Roth and Bourgine (2004) [18] – define an epistemic community by the group of people that maximally share a number of concepts). These clusters are not strictly 'Communities of Practice' (Lave & Wenger, 1991 [19]) because the authors need not be acquainted or involved in a common practical task. Regardless, a community of practice is certainly a special type of knowledge community. The participation in Wikipedia as a whole, although a theme to be developed elsewhere, can be said to be a large community of practice where editors interact using shared paradigms, meanings, values and practices and where a lot of the learning is tacit: wikipedians learn how to edit articles, how to fight vandals, how to use policies to make their points through, how to present themselves in user pages and so on.

Bryant, Forte and Bruckman (2005) [20] in 'Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia," argued that, "observations of members' behavior in Wikipedia reveals that the three characteristics of Communities of Practice identified by Wenger are strongly present on the site: community members are mutually engaged, they actively negotiate the nature of the encyclopedia-building enterprise, and they have collected a repertoire of shared, negotiable resources including the Wikipedia software and content itself."

Clustering allows us to zoom in into this community of practice and detect more specific modules bounded by shared interest. In WikiProjects, in particular, authors are involved in a common task and are, therefore, creating structures that are closer to mini-communities of practice than the other 'epistemic communities'.

Finally, it should be noted that in order to complete the typology with the clusters found in the data, some of the clusters contain a number of articles in topics as diverse as 'Tsunami' and 'Albert Einstein'. It is not surprising that this module is diverse. The $K_{4,1}$ cliques are 4 editors that have co-edited just one article. If they had co-edited two or more articles, then one would expect more similarity (in general the articles become more homogeneous). This type of clique with a low second index means that the articles do not have anything in common. This type of clusters are a hint that Wikipedia is also a product of more loose dedications by people who edit in articles which are more broad, but also that there are different editing patterns, and not all are content-driven. Editing to fix typos, or to make tables of contents can also group people.

As this is a study with both quantitative and qualitative features, we used semantic categories top-down to describe the kinds of clusters found in the data, harvested bottom-up. This way we assessed qualitatively the nature of collaboration between editors in a subset of the English Wikipedia, grounded on network analysis.

5. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

A more systematic study could reveal possible 'network signatures', *i.e.* ways to identify controversies or WikiProjects directly from the network structure. It will be interesting to complement the insights discovered from our meso-level modules with a deeper probe into the specific connections through discussion pages, on the level of individual articles and paragraphs to understand the patterns of distributed work and perhaps, cognition in greater depth. Analyzing the network of modules informs us about the individual modules and their structural relation to each other. Further work could provide information to understand the network of Wikipedia in relation to other networks (scientific collaborations, open source projects). In the future, it will be interesting to expand the bipartite clustering technique and approach to other areas of Wikipedia and other datasets; to organize the algorithm in order

to allow for the module surrounding any article to be visualized, and contextualize the findings in the light of more abstract claims of the power of technology and cluster, in specific wikis and wikipedias to organize knowledge, work together, and ultimately be part of a cognitive system that comprises humans, technologies and values. It will also be interesting to pursue the interdisciplinary meso-level of analysis as it seems to result in insights which inhabit the area between the quantitative patterns and the qualitative details usually found by other more traditional disciplines.

6. CONCLUSION

Detecting modules of articles and editors in Wikipedia yields important insights into the nature of collaboration. The technique used in the present research probes a level where collaboration is surely taking place because people in fact gather around a number of articles and work intensely on them.

7. ACKNOWLEDGMENTS

Rut Jesus acknowledges support by the Portuguese Foundation for Science and Technology with the grant SFRH/BD/ 27694/2006 and would like to thank Camille Roth for discussions and bibliography concerning biclique history. Sune Lehmann acknowledges support by the Danish Natural Science Research Council and James S. McDonnell Foundation 21st Century Initiative in Studying Complex Systems, the National Science Foundation within the DDDAS (CNS-0540348), ITR (DMR-0426737) and IIS-0513650 programs, as well as by the U.S. Office of Naval Research Award N0001407-C and the NAP Project sponsored by the National Office for Research and Technology (KCKHA005).

8. REFERENCES

[1] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: the case of wikipedia," arXiv:physics/0602026, Feb 2006.

[2] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi, "Temporal analysis of the wikigraph," in WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, (Washington, DC, USA), pp. 45–51, IEEE Computer Society, 2006.

[3] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet, "Wikipedias: Collaborative webbased encyclopedias as complex networks," arXiv:physics/0602149, Jul 2006.

[4] F. Ortega and J. M. Gonzalez-Barahona, "Quantitative analysis of the wikipedia cluster of users," in WikiSym '07: Proceedings of the 2007 international symposium on Wikis, (New York, NY, USA), pp. 75–86, ACM, 2007.

[5] A. Capocci, F. Rao, and G. Caldarelli, "Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia wikipedia," EPL (Europhysics Letters), vol. 81, no. 2, pp. 28006+, 2008.

[6] F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham, "Talk before you type: Coordination in wikipedia," in System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on, pp. 78–78, 2007.

[7] D. M. Wilkinson and B. A. Huberman, "Assessing the value of coooperation in wikipedia," arXiv:cs/0702140, Feb 2007.

[8] A. Kittur, and R.E. Kraut, "Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination." CSCW 2008: Proceedings of the ACM Conference on Computer-Supported Cooperative Work. New York: ACM Press, 2008.

[9] F. B. Viegas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in CHI '04: Proceedings of the 2004 conference on Human factors in computing systems, pp. 575–582, ACM Press, 2004.

[10] R. P. Biuk-Aghai, "Visualizing co-authorship networks in online wikipedia," in Communications and Information Technologies, 2006. ISCIT '06. International Symposium on, pp. 737–742, 2006.

[11] B. Suh, E. H. Chi, B. A. Pendleton, and A. Kittur, "Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations," in Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on, pp. 163–170, 2007.

[12] Camille Roth, "Co-evolution in Epistemic Networks – Reconstructing Social Complex Systems", Structure and Dynamics: eJournal of Anthropological and Related Sciences : Vol. 1: No. 3, Article 2, 2006.

[13] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature 435, pp. 814-818, 2005.

[14] S. Lehmann, M. Schwartz, and L. K. Hansen, "Biclique communities," arXiv:0710.4867, 2007.

[15] A. Halavais and D. Lackaff, "An analysis of topical coverage of wikipedia," Journal of ComputerMediated Communication, vol. 13, no. 2, pp. 429– 440, 2008.

[16] F. B. Viégas, M. Wattenberg, and M. Mckeon, "The hidden order of wikipedia," Online Communities and Social Computing, pp. 445–454, 2007.

[17] R. Singel, "Wikipedia Bans Church of Scientology" in Wired.com, May 29, 2009. Retrieved on July 4th, 2009 from http://www.wired.com/epicenter/2009/05/wikipedia-bans-church-of-scientology.

[18] C. Roth and P. Bourgine, "Epistemic communities: description and hierarchic categorization," arXiv:nlin/0409013, Sep 2004.

[19] J. Lave and E. Wenger, *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, 1991.

[20] S. Bryant, A. Forte and A. Bruckman, "Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia," Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, November 06-09, 2005.