6-2008

WikiNetViz: Visualizing friends and adversaries in implicit social networks

Minh-Tam LE
Nanyang Technological University

Hoang-Vu DANG

Ee Peng LIM
Singapore Management University, eplim@smu.edu.sg

Anwitaman DATTA
Nanyang Technological University

DOI: https://doi.org/10.1109/ISI.2008.4565029

WikiNetViz: Visualizing friends and adversaries in implicit social networks

Conference Paper · July 2008
DOI: 10.1109/ISI.2008.4565029 · Source: IEEE Xplore

4 authors, including:

Hoang-Vu Dang
University of Illinois, Urbana-Champaign
19 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Ee-Peng Lim
Singapore Management University
448 PUBLICATIONS **9,087** CITATIONS
SEE PROFILE

WikiNetViz: Visualizing Friends and Adversaries in Implicit Social Networks

Minh-Tam Le, Hoang-Vu Dang, Ee-Peng Lim, Anwitaman Datta
Centre for Advanced Information Systems
School of Computer Engineering, Nanyang Technological University,Singapore 639798
Email: {MinhTamLe,DANG0010,aseplim,Anwitaman}@ntu.edu.sg

***Abstract*—When multiple users with diverse backgrounds and beliefs edit Wikipedia together, disputes often arise due to disagreements among the users. In this paper, we introduce a novel visualization tool known as WikiNetViz to visualize and analyze disputes among users in a dispute-induced social network. WikiNetViz is designed to quantify the degree of dispute between a pair of users using the article history. Each user (and article) is also assigned a controversy score by our proposed *ControversyRank* model so as to measure the degree of controversy of a user (and an article) by the amount of disputes between the user (article) and other users in articles of varying degrees of controversy. On the constructed social network, WikiNetViz can perform clustering so as to visualize the dynamics of disputes at the user group level. It also provides an article viewer for examining an article revision so as to determine the article content modified by different users.**

***Index Terms*—Controversy, Disputes, Visual analytics, Wikipedia.**

I. Introduction

A. Motivation

Wikipedia, as an open collaborative encyclopedia, embraces consensus building as one of its founding policies.[1] The evolving content of Wikipedia articles relies heavily on the collaboration among contributors. Throughout the history of an article, contributors edit content by adding and deleting each other's contribution. One can also find a mixture of concordance and conflict, collaboration and negotiation among the contributors [1].

To manage the amount of conflicts caused by co-editing, Wikipedia has introduced not only guidelines for consensus-building, but also policies against disruptive editing [2] and ownership of articles.[3] These guidelines and policies however can offer only limited help as they cannot guarantee all contributors will work harmoniously together at all times. The reality is that many contributors, due to their diverse backgrounds and beliefs, exert different opinions in the article content causing disputes of varied degrees. Even with moderation, such differing views may not be easily resolved, and may eventually degrade the quality of the Wikipedia content.

Disputes is closely linked to controversial topics in Wikipedia. When an article involves some controversial topics, it is more likely that it will invite disputes. On the other hand, we also expect some contributors to be controversial when they easily initiate disputes with others whether or not the topics involved are controversial. With this above associations between inter-contributor disputes, article controversy and contributor controversy, it is therefore natural to study them all together.

In Wikipedia, unfortunately, it is difficult to detect disputes, controversial articles and controversial contributors due to the massive amount of article content and history data. While different models and techniques can be developed, it is still vital to have a range of visualization tools to help analyzing them.

B. Research Objectives and Contribution

In this paper, we therefore propose to address the visualization problem for disputes among contributors and the articles in dispute. We believe that a visualization tool is required to quickly identify the controversial articles that are likely to contain disputes, and to easily view the amount of disputes among contributors so as to determine the nature of disputes, the topics involved, and to resolve the dispute if possible.

Our research objectives and contributions to address the problem are as follows:

- We have developed a model for measuring the extent of dispute between two contributors by examining the amount of word deletion among contributors. This further allows us to derive a dispute-induced social network among contributors. Such a social network can be easily presented in a visual tool for analysis purposes.
- To measure the degree of controversy associated with article topics and contributors, we introduce the **ControversyRank** model that can automatically assign a controversy score to each Wikipedia article (contributor) based on the amount of disputes between contributors in different articles. These controversy scores will help us to identify controversial articles for detecting disputes and to differentiate controversial contributors from non-controversial ones during our visualization.
- We have developed **WikiNetViz**, a visualization tool, using the above dispute definitions and ControversyRank model. With WikiNetViz, the dispute-induced social network can be easily visualized and the groups of contributors in dispute can also be found using a clustering algorithm. This will allow the disputes among contributors

[1]http://en.wikipedia.org/wiki/WP:CON
[2]http://en.wikipedia.org/wiki/WP:DISRUPT
[3]http://en.wikipedia.org/wiki/WP:OWN

to be highlighted for attention.
- We will illustrate the use of WikiNetViz using an example article which is known to be controversial. We show that by using WikiNetViz, the disputes in the article can be shown clearly and one can determine the conflicting contributor groups with good accuracy.

C. Paper Outline

For the rest of this paper, Section II overviews the related work. We will present our dispute definition and dispute-induced social network in Section III. WikiNetViz will be described in Section IV followed by a case study discussion in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

Various works have been carried out on collaborative social networks, most focusing explicitly on Wikipedia. The research interests were mainly on examining article quality [2], [3], [4], contributor reputation and trustworthiness [5], [6], [7], and article content evolution [8], [1]. Among his many data-visualization projects, Chris Harrison developed WikiViz [9] which displays the complex, dynamic relationships among millions of Wikipedia topics.

As presented by Stvilia et al. [3], Wikipedia has dedicated a significant portion of its massive effort in coordinating its *information quality* (IQ) assurance work. A part of the effort goes to managing conflicts and disputes among Wikipedians. Interestingly, the paper pointed out the existence of formally declared philosophies on quality assurance "*through establishing associations [which] may help to make [Wikipedia] more predictable*". This fact reveals a more complex structure within the Wikipedia community, where rivalry and cooperation are the driving force behind quality.

Sabel [8] proposed the *adoption coefficient*, which indicates the similarity between two corresponding article revisions, to build a tree structure which reflects the evolution of the article, together with the editing activities among contributors. With the same focus, Viegas et al. [1] visualized how article contents evolved through edit histories with their *history flow* tool, highlighting patterns of contributors' edit behaviors.

Orthogonal to the *history flow* approach which focuses on content evolution, Kittur et al. [10] built *Revert Graph* to discover conflicts among Wikipedians. They also proposed a supervised classification method to automatically identify controversial articles.

Brandes and Lerner [11] offered a more general approach to analyze disagreements among Wikipedia contributors by constructing the revision network, since reverts are not the only and the best indicators of conflicts. They also proposed a spectral layout method to visualize conflicts among contributors. However, the revision network, designed based on the idea of identifying controversy in the reply-to network among Newsgroup users, may not be suitable, as Wikipedia articles' revisions, recorded in chronological order, does not strictly represent a reply-to relationship.

Complimentary to the supervised classification approach from [10], in our earlier work [12], we have also proposed the ControversyRank model, which examines edit histories to identify controversial articles and contributors. Similar to the idea of the reply-to relationship in the revision network, our model scans the revision histories, and records the words which one contributor deleted from another as indicators of conflicts. Instead of relying on statistical metrics as in Kittur et al., by going through the revision history, our model analyze the evolution of article content, as in the *history flow* visualization, with an additional focus on the article-contributor and contributor-contributor relationship.

III. Disputes in Controversial Articles

Disputes among contributors in Wikipedia can be observed from the historical information of the articles. Most of the disputes happen for at least one of the two reasons: (a) the controversial nature of some particular topics, and (b) the combative editing behavior of some contributors.

In this section, we first introduce a way to measure disputes in Wikipedia articles. This leads us to develop an implicit social network induced by disputes among contributors. As disputes often occur in controversial topics, we will describe our earlier proposed model (i.e., CR Model) for identifying articles that cover controversial topics.

A. Disputes and Dispute-Induced Social Networks

In Wikipedia, disputes occur but are not well captured in the database. Contributors can tag an article or a section of an article to be in dispute or controversial but disputes between contributors are hardly recorded. Sometimes, one can find comments left by contributors in the discussion page of an article mentioning disputes between contributors. Such text comments, however, are not always machine understandable.

Hence, in our paper, we model a dispute between contributors u_i and u_j by the deletion of u_j's content by u_i, and deletion of u_i's content by u_j. Given that there are many articles and contributors, disputes can therefore be represented by a bipartite graph as shown in Figure 1. The graph consists of a set of contributor ordered pairs u_i, u_j's and a set of articles r_k's. Each directed edge from (u_i, u_j) to r_k has a weight d_{ijk} indicating the amount of disputes. We measure the amount of disputes by the number of u_j's contributed words in r_k which were removed by u_i in article r_k.

Note that disputes can be *active* or *passive*. From a user u_i's standpoint, active dispute with another user u_j in an article r_k is represented by d_{ijk}. The passive dispute with u_j in r_k is represented by d_{jik}. In most cases, $d_{ijk} \neq d_{jik}$.

From the above bipartite graph, we derive the **dispute-induced social network** for a set of article R which is defined by a set of nodes representing contributors and edges between nodes representing disputes. An edge between u_i and u_j is assigned the weight $\sum_{r_k \in R}(d_{ijk} + d_{jik})$, i.e., the total amount of dispute between u_i and u_j in articles from R. When R consists of only one article, the dispute-induced social

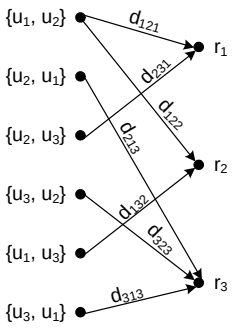

Fig. 1. Articles and contributors disputes represented by a bipartite graph.

network represents disputes among contributors in the article (see Figure 2).

Dispute-induced social network is an *implicit* social network as its edges do not suggest proximity among members of the network. On the contrary, the edges carry negative relationships. When no edge exists between two contributors, it just simply means that there are no disputes detected between them. Beyond that, we are not able to immediately conclude that they enjoy positive relationship.

B. ControversyRank Model and Controversial Articles

Recently, we proposed the **ControversyRank (CR)** Model [12], which identifies controversial articles in Wikipedia based on the disputes mentioned in Section III-A.

The CR model exploits a *mutual reinforcement relationship* [13] between articles and contributors. It seeks to use this mutual reinforcement relatonship to determine both controversial articles and controversial contributors. The mutual reinforcement relationship can be summarized by the following presumptions:

- An article is more controversial if it attracts more disputes among less controversial contributors.
- Likewise, a contributor is deemed to be more controversial if she is involved in more conflicts in less controversial articles.

The model assigns a controversy score to each article and contributor using Equations 1 and 2 respectively.

$$C_k^r = \frac{\sum_{i,j} agg[(1-C_i^u),(1-C_j^u)] \times d_{ijk}}{\sum_i o_{ik}} \quad (1)$$

$$C_i^u = \frac{\sum_{j,k}(1-C_k^r) \times (d_{ijk}+d_{jik})}{\sum_{j,k} o_{jk} \times I(i,j,k) + \sum_k o_{ik}} \quad (2)$$

where o_{ik} represents the number of words contributed by u_i to r_k, and $I(i,j,k))$ is a boolean function which indicates whether u_i has deleted any word from u_j in r_k.

The article controversy score in Equation 1 is taken for the sum of disputes in the article weighted by the aggregated inverse controversy of the conflicting pairs of contributors. The aggregate function agg can take either the average or the product of the inverse controversy scores of the two contributors involved. The contributor controversy score in Equation 2 is computed in a similar manner by summing all the disputes engaged by u_i, weighted by the inverse article controversy scores. More details about the CR model and its accuracy performance can be found in [12].

IV. WikiNetViz and Implicit Contributor Social Network

Based on our proposed dispute measures and ControversyRank model, we built a visualization tool called **WikiNetViz** to analyze contributor conflicts in a selected article, and the relationships among the contributors.

A. Dataset

To analyze Wikipedia articles containing disputes, we constructed an article dataset by gathering articles from the *Science* category of *English Wikipedia*. A list of 37,489 article titles were crawled from the *Science* category. The crawler was configured to gather article titles from only the top 3 levels of the category hierarchy.[4] Based on this list of titles, 25,571 articles together with their edit histories, were found and extracted from the Wikipedia database dump created in November, 2006. There are 310,287 distinct contributors in this dataset.

For each article, we derived the amount of disputes among the contributors by examining each pair of successive revisions $(r(t), r(t+1))$. After removing stop words from the two revisions, we applied a comparison algorithm [14] on them, identifying matching words and differing ones. Words in $r(t)$ which are absent from $r(t+1)$ are considered deleted by the author of $r(t+1)$. Similarly, words in $r(t+1)$ for which there are no correspondences in $r(t)$ are counted toward the contribution by $r(t+1)$'s author. In this process, we considered only pairs of successive revisions which were authored by different contributors. In a succession of revisions made by the same user, we took into account only the last revision, ignoring all intermediate ones. The resulting differences between the revisions were used to derive the dispute-induced social network.

B. Visualizing Dispute-Induced Social Networks in WikiNetViz

WikiNetViz supports visualization of the dispute-induced social network of a selected article as shown in Figure 2. Since articles are assigned controversy scores using the CR Model, the user can choose to view an article among the controversial ones. On the lower left corner of the interface, the user can choose to display only those edges which exceeds a specified minimum weight w, or to display k contributors with the highest number of disputed words in the article. k is assigned 10 by default. Figure 2 shows the 30 contributors with highest number disputed words from the Wikipedia article *Ars Technica* [5].

Each contributor is a node in the network. The height of the node is proportional to the contributor's controversial score

[4]Wikipedia maintains a hierarchical structure of article topics. There are many levels of sub-categories under the *Science* category

[5]http://en.wikipedia.org/wiki/Ars_technica

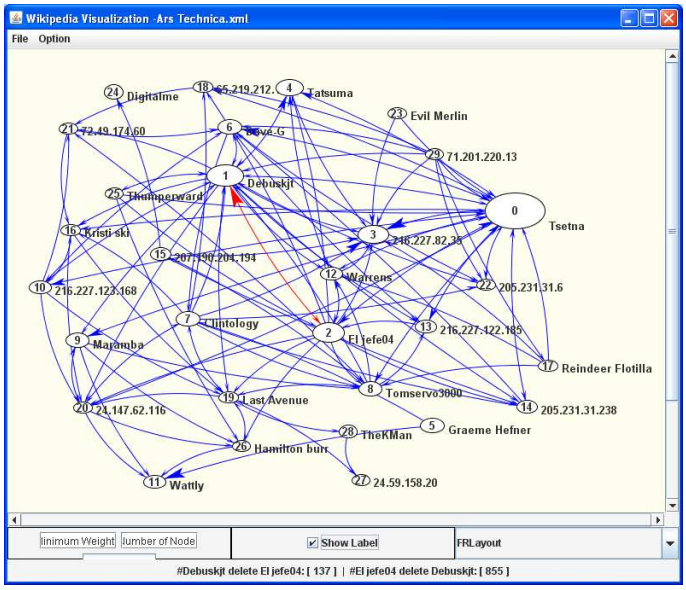

Fig. 2. Dispute-induced social network of *Ars Technica* article

computed by the CR model using the constructed dataset, indicating how controversial the contributor is. The width of the node is proportional to the number of words deleted by that particular contributor in the selected article. The number inside the node indicates the contributor's rank order in decreasing contributor's controversy scores. Checking the "Show Label" checkbox allows the usernames of contributors to be displayed. Different layout options are also available at the lower right corner of the interface.

The size of the arrow head at the end of an edge (u_i, u_j) is proportional to the number of words u_i removed from u_j in the article. Upon clicking an edge (u_i, u_j), the tool will display at the bottom of the screen the conflict information between u_i and u_j in the article. By double-clicking on a contributor node, a pop-up window will appear showing the contributor's username, global controversial score, as well as the number of words she (actively) deleted from others, and the number of her words (passively) removed by others.

C. Clustering of Controversial Contributors

In exploring the relationships among the involved authors, it is interesting not only to know who is in dispute with whom individually and how much the intensity of such pair-wise dispute is, but also to discover groups of contributors which are in dispute. Such disputing group information can help the user to understand the group dynamics among contributors. Contributors of the same group, on the other hand, tend to have little dispute with one another. They may possess similar opinions on the article subject matter or just have not had disputes with one another.

To achieve the above, we conduct clustering on the dispute-induced social network. Unlike the traditional clustering problem, our social network contains edges representing disputes instead of similarities. We therefore derive the similarity as follows.

We first derive D_{ij}, the total dispute between u_i and u_j defined by the number of disputed words between users u_i and u_j in the article.

$$D_{ij} = d_{ij} + d_{ji} \tag{3}$$

The similarity measure between u_i and u_j can then be derived as

$$S_{ij} = \max_{k,l} D_{kl} - D_{ij} \tag{4}$$

We applied the contributor's similarity matrix obtained by Equation 4 to the CLUTO clustering toolkit [15] to divide the contributors into different clusters. There are two available clustering methods (hierarchical and min-cut graph) which the user can select from a pop-up menu. In addition, the user is able to specify the desired number of clusters by moving the scroll bar on the same menu.[6]

D. Visualization of article content

The user can view the Wikitext[7] content of an article in the content frame of WikiNetViz. Once a particular revision of an article is selected from the drop-down box, the revision content will be shown in the content frame with the name of the authoring contributor shown in the title of the frame. By checking the "Deleted" and "New" options beside the Revision selector, the user will be able to highlight which part of the revision content is newly inserted and which part would be deleted in a later revision. If the user clicks on a contributor node of the dispute-induced social network, that contributor's contribution in the selected revision (if any) will then be highlighted. Moving the mouse pointer over a particular word will reveal the author of that word, the timestamp (and revision number) of insertion, as well as the remover and time of removal (in case the word would be deleted in a later revision).

Figure 3 shows the visualization of *Ars Technica*'s 405^{th} revision when a contributor *Maramba* is selected in the network view. This revision was authored by contributor *Debuskjt* as shown in the window frame title. The words highlighted in cyan are those contributed by the contributor *Maramba* that still remained in the 405^{th} revision of the article. The words highlighted in yellow are those which would be deleted in a later revision.

V. Case Study

In this case study, we use WikiNetViz to analyze the dispute-induced contributor network of the article *Ars Technica*, an article listed in Wikipedia's list of controversial topics[8] and ranked by our CR model as the 4th most controversial article in the data set. The history of the article until November, 2006 (which is the time of the dataset used in our experiments)

[6]WikiNetViz allows 2 to 8 way clustering. Although more than 8 clusters partitioning is possible, we believe that 8 is sufficient for most practical purposes

[7]Wikitext is the markup language used in Wikipedia.

[8]http://en.wikipedia.org/wiki/Category:Wikipedia_controversial_topics

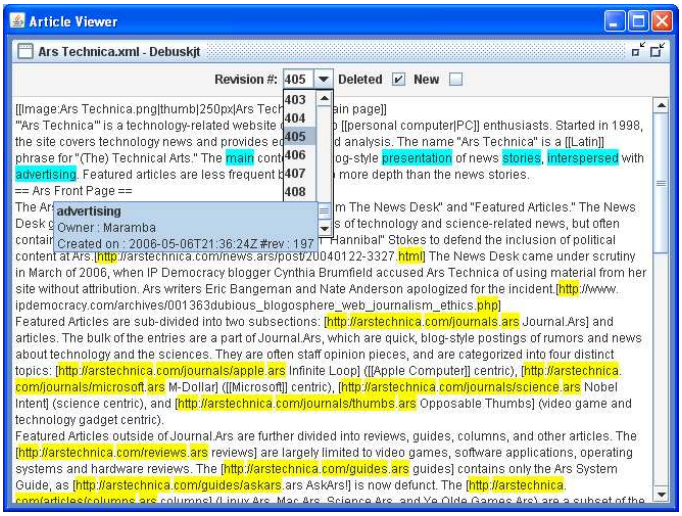

Fig. 3. Visualization of 405^{th} revision of the *Ars Technica*.

TABLE II
CONTRIBUTOR GROUPS IN *Ars Technica* ARTICLE.

Group	Supporting	Opposing	Neutral	Total
A	8	0	3	11
B	3	0	1	4
C	0	8	1	9
D	2	3	1	6

has recorded a bitter controversy lasting from February 20^{th} till September 8^{th} 2006 and involved 36 contributors, with several attempts for negotiation and meditation. The article refers to a technology-related website.[9] As reflected in the edit history of the article, 19 contributors insisted on including several criticisms about the site, while others opposed the idea, saying that the criticisms were unjustifiable.

Going through the discussion page and archives, we could further decompose the opposition group into 3 subgroups: (i) those who immediately removed all criticisms, (ii) those who negotiated for a separate criticism section with verified content, and (iii) those who negotiated to blend verified criticism into various parts of the article. The usernames of the contributors involved in the controversy are listed in Table I according to their opinion groups.

Using the clustering module of WikiNetViz, we partition the top-30 contributors of *Ars Technica* into 4 groups (A to D), as shown in Figure 4. It is obvious that our 4-way clustering result does not match the actual partitioning in Table I. The reason can be that there were not many disputed words among the 3 subgroups of the opposition camp and 2 of the 3 subgroups consist of only 2 contributors each.

For verification purpose, we manually classified the contributors as "*Supporting*" or "*Opposing*" or "*Neutral*",[10] based on his/her stand on the dispute and compared our labels against the clustering results, as shown in Table II.

[9]http://www.arstechnica.com

[10]Note that the "*Neutral*" label is needed since some of the top-k contributors may not necessarily be involved in the dispute

As it turned out, 24 out of 30 contributors were involved in the dispute. Excluding the 6 neutral contributors, groups A and B in our cluster result consist of the supporters of the criticism section, while all contributors from groups C and D (except contributors *Dave-G* and *El jefe04* in group D) are part of opposition. Furthermore. another interesting point to note is that the non-neutral contributors in groups B and C are those who fervently fought over the issue throughout the period of the controversy. Hence, we are able to conclude that the clustering results closely represent the "*Supporting* vs. *Opposing*" opinion groups. The existence of neutral contributors and the misclassification of disputing contributors can be attributed to the fact that our approach take all the deleted content, ignoring the semantic context, and therefore may have missed important details necessary to refine the ranking and clustering.

VI. CONCLUSION

Given Wikipedia's large number of articles and contributors, it is challenging to visualize disputes among contributors, disputes among groups of contributors, and using the information to analyze controversial articles and contributors. In this paper, we determine disputes from the article history and using them to induce an implicit social network. By applying disputes to a ControversyRank model, the controversy scores of articles and contributors can be computed. WikiNetViz, a visualization tool, has been developed to visualize contributors' disputes and to cluster them into contributor groups. A case study using WikiNetViz to analyze a controversial article has been illustrated.

In contrast to explicit social networks where there are obvious positive association among nodes, an implicit network uses other semantics to determine node associations. In the context of Wikipedia, these implicit relations are antagonistic in nature. From the basic deletion based implicit network, group dynamics are not obvious. However, once a clustering is done, putting members with common antagonists in groups, we can also identify "like-minded" members getting clustered together. Such clustering of users within a single document is not necessarily a hard evidence of positive association among the members, nevertheless, it is a reasonably good indicator. Such groups are a clear indication of lobbies for corresponding belief or interpretation over which the disputes occur. It also indicates that disputes are often among these groups rather than among only individuals. We hope to use this insight to derive a better controversy ranking of articles by augmenting the group based conflicts with the individual disagreements in the mutual reinforcement relationships. Also, currently, once the clustering is completed, we omit the fact that some separate clusters are closer to each other, and are relatively more antagonistic to other clusters. Ignoring such inter-cluster affinity is one of the reasons why we did not obtain a perfect match in our case study between the computed cluster versus the manually generated groups. Addressing these issues are some of our immediate next steps.

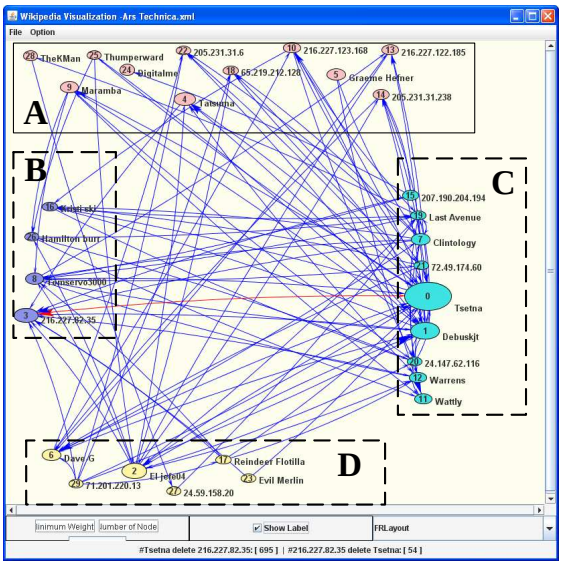

Fig. 4. Four-way clustering of contributors in *Ars Technica* article.

TABLE I
CONTRIBUTOR GROUPS IN *Ars Technica* ARTICLE.

Opinion	Contributor list
Supporting	216.227.56.73, Kristi, Tomervo3000, 205.231.146.195, 216.227.123.168, 24.105.219.78, Maramba, Dave-G, 216.227.82.35, 205.231.31.238, 205.231.31.6, 65.219.212.128, 216.227.83.118, 205.231.151.88, 216.227.122.185,67.123.205.241, Digitalme, Tatsuma, El jefe04
Outright opposing	157.91.44.1, 155.33.109.95, DrPizza, Clintology, On-no, 24.147.62.116,65.161.188.11, 72.49.174.60, Evil Merlin, 207.190.204.194, 71.201.220.13, FlyPenFly, Last Avenue
Opposing - negotiating for a separate, verified criticism section	Tsetna, Reindeer Flotilla
Opposing - blending verified criticism into other sections	Debuskjt, Warrens

ACKNOWLEDGMENTS

This work was supported in part by A*STAR Public Sector R&D, Project Number 062 101 0031. We thank our anonymous reviewers for their comments and suggestions.

REFERENCES

[1] F. B. Viegas, M. Wattenberg, and K. Dave, "Studying Cooperation and Conflict between Authors with history flow Visualizations," in *Proceedings of CHI'04*, April 2004.

[2] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong, "Measuring Article Quality in Wikipedia: Models and Evaluation," in *Proceedings of CIKM'07*, November 2007.

[3] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Information Quality Work Organization in Wikipedia," *Journal of American Society for Information Science and Technology*, accepted 2007.

[4] R. Priedhorsky, J. Chen, S. T. K. Lam, katherine Panciera, L. Terveen, and J. Riedl, "Creating, Destroying, and Restoring Value in Wikipedia," in *Proceedings of GROUP'07*, November 2007.

[5] B. T. Adler and L. de Alfaro, "A Content-Driven Reputation System for the Wikipedia," in *Proceedings of WWW2007*, May 2007.

[6] D. L. McGuinness, H. Zeng, P. P. da Silva, L. Ding, D. narayanan, and mayukh Bhaowal, "Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study," in *Proceedings of WWW2006*, May 2006.

[7] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness, "Computing Trust from Revision History," in *Proceedings of International Conference on Privacy, Security and Trust*, October 2006.

[8] M. Sabel, "Structuring Wiki Revision History," in *Proceedings of the 2007 international symposium on Wikis*, October 2007.

[9] C. Harrison, "Wikiviz: Visualizing wikipedia." [Online]. Available: http://www.chrisharrison.net/projects/wikiviz/index.html

[10] A. Kittur, B. Suh, B. A. Pendleton, and E. h. Chi, "He Says, She Says: Conflict and Coordination in Wikipedia," in *Proceedings of CHI'07*, April 2007.

[11] U. Brandes and J. Lerner, "Visual Analysis of Controversy in User-generated Encyclopedias," in *IEEE Symposium on Visual Analytics Science and Technology*, October 2007.

[12] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, H. W. Lauw, and K. Chang, "On Ranking Controversies in Wikipedia: Models and Evaluation," in *Proceedings of the WSDM'08*, February 2008.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, September 1999.

[14] P. Heckel, "A Technique for Isolating Differences Between Files," *Communications of the ACM*, vol. 21, no. 4, pp. 199–222, April 1978.

[15] G. Karypis, "CLUTO - Software for Clustering High-Dimensional Datasets ." [Online]. Available: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview