Research Repository UCD

Title	Integration of multiple network views in Wikipedia
Authors(s)	Wu, Guangyu, Cunningham, Pádraig
Publication date	2015
Publication information	Wu, Guangyu, and Pádraig Cunningham. "Integration of Multiple Network Views in Wikipedia." Springer, 2015. https://doi.org/10.1007/s10115-014-0802-7.
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/6416
Publisher's statement	The final publication is available at www.springerlink.com
Publisher's version (DOI)	10.1007/s10115-014-0802-7

Downloaded 2026-05-01 23:36:19

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

Integration of Multiple Network Views in Wikipedia

Guangyu Wu, Pádraig Cunningham
School of Computer Science & Informatics,
University College Dublin, Ireland
{guangyu.wu,padraig.cunningham}@ucd.ie

Abstract

One of the challenges in network data analysis is the determination of the most informative perspective on the network to use in analysis. This is particularly an issue when the network is dynamic and is defined by events that occur over time. We present an example of such a scenario in the analysis of edit networks in Wikipedia – the networks of editors interacting on Wikipedia pages. We propose the prediction of article quality as a task that allows us to quantify the informativeness of alternative network views. We present three fundamentally different views on the data that attempt to capture structural and temporal aspects of the edit networks. We demonstrate that each view captures information that is unique to that view and propose a strategy for integrating the different sources of information.

1 Introduction

Insight into objects embedded in a network of interactions can be derived from an analysis of the structure of that network. This fundamental idea has been demonstrated not only in social network analysis [14, 21] and in bioinformatics [23] but also in areas as diverse as spam filtering [6, 5], telecommunications [3], fraud detection [24] and of course web search [16]. How best to represent the network around an object is still a significant research challenge. Recently profiling using network motif counts has emerged as a promising solution for characterizing networks [23, 26, 27, 3].

In this paper we analyse edit networks in Wikipedia and consider three alternative views on the data that are significantly different. We evaluate the effectiveness of motif-based characterisations of networks in these three views in terms of the effectiveness for predicting article quality. We also evaluate strategies for integrating these three representations in order to improve predictive accuracy.

The three views capturing temporal, structural and edit trajectory information are as follows:

- The **temporal network** view considers article revision history as a sequence of editor's interactions (add, delete, revert) and the motifs are the various interaction combinations (see Figure 2).
- The **ego network** is a static bipartite network where major edits are represented as edges between article and editor nodes (see Figure 3).
- The **trajectory network** represents article revisions through a sequence of editors (see Figure 6).

Using each of these three views an article can be characterized by vectors of motif counts (vectors of length 125, 22 and 13 respectively). The objective of our study is to assess the effectiveness of these feature-based representations for predicting article quality using supervised machine learning methods.

In the next section we briefly review some relevant research on social network analysis and authoritativeness. We describe the three network views (temporal network, structural network and

trajectory network) in Sections 3, 4 and 5, respectively and follow by showing their performance on three Wikipedia article collections in Section 6. The two integration strategies are introduced in Section 7 and we demonstrate the improvement resulting from integrating the three network views in Section 8. The final section 9 concludes with a summary of the integration work.

2 Related Research

The ultimate objective of the network analysis we propose here is to assess the quality and authoritativeness of user generated content (UGC) in the social web. UGC has been enormously disruptive as resources such as Wikipedia, TripAdvisor and MenuPages replace traditional businesses that provided similar services. At the same time, this raises questions regarding the quality of UGC [4]. Is an answer on *Yahoo! Answers* correct? Are details presented in blogs or on Wikipedia credible? Can we trust reviews on TripAdvisor or on Amazon? These questions have received a good deal of attention in recent research. We focus on work related to Wikipedia here.

In contrast to traditional encyclopedias, where authority derives from expert contributors, Wikipedia depends on collaboration and consensus to produce quality articles. There has been some controversial research that suggests that the quality of Wikipedia articles approaches that of established encyclopedias [9]. The famous quote from Surowiecki's *The Wisdom of Crowds* is that "under the right circumstances, groups are remarkably intelligent" [25]. The challenge when using Wikipedia is to determine whether the circumstances that produced the article in question were *right* or not.

The first requirement for a Wikipedia article to be authoritative is the *many eyes* idea [19] – a significant group of contributors must have cooperated to produce the article. It is also important that this collaboration has been constructive and it is better if the editors have a reasonable reputation as contributors. Adler and De Alfaro [1] have pursued a content driven strategy to assess editor reputation. They have used text survival and edit distance to quantify editor reputation. In later work [2] they show that edit longevity is a good measure of editor contribution.

Korfiatis et al. [17] pursue a network-based strategy to evaluate authoritative sources in Wikipedia. They construct a two-mode network of articles and contributors. The article nodes are linked by hyperlinks and contributors are linked if they have worked on the same article. Contributors are also linked to articles on which they worked. The study proposes article and contributor degree centrality as indicators of authoritativeness. This is similar in spirit to the strategy in our work as degree centrality is captured by a subset of the network motifs we consider.

Brandes et al. [7] have also analysed the collaboration structure in Wikipedia. Their work has focused on the edit interactions on individual articles. Edges between individual contributors represent *delete*, *undelete* and *restore* interactions. The main contribution of this work is to present the notion of bipolarity that captures the level of conflict between the contributors to an article. Thus the work is more directed at the problem of Wikipedia vandalism than the issue of authoritativeness that is the subject of this paper.

Laniado et al. [18] presented an algorithm that assigns scores to all contributors of a Wikipedia article according to their contribution, and selects the top contributors to build a collaboration network of authors where edges represent the co-authorship between authors. Thus the inexperienced authors are filtered out and the co-authorship networks becomes more informative. With the exception of eigenvector centrality (where edge weights were considered) the features they extracted were taken from unweighted versions of the networks.

Recently, Jurgens et al. [13] proposed temporal motifs extracted from temporal bipartite networks using Wikipedia article revision histories. In the work they defined editor interactions as *minor add*, *major add*, *minor edit*, *major edit*, *minor delete*, *major delete* and *revert*. A large set of temporal motifs was built of these interactions. They employed these motifs to classify pages as combative or cooperative and in a further study they analysed Wikipedia's content growth with these temporal motifs.

Related to this temporal network representation is the idea of a trajectory network [12, 15]. Iba et al. [12] introduce the idea of a trajectory network for a Wikipedia article as the network of editors the article has passed though during its development (see Section 5 for more detail). They show that this representation helps highlight different editor roles including *coolfarmers*, the outstanding

editors that coordinate the editing process, and *egoboosters* who are not so cool. Keegan et al. [15] go on to show how this representation is effective for revealing characteristic patterns in breaking news collaborations in Wikipedia.

In earlier work we have used a static bipartite network representation of edit activity around a Wikipedia article as an indicator of the quality of the article [26, 27] Articles are classified using a set of network motifs and motif count ratios. We have identified a subset of motifs that are predictive of Wikipedia quality as determined by the official Wikipedia quality scale that ranges from *Stub* and *Start* articles at the bottom end to *Featured* articles at the top.

There also exists non-network based studies, for example, Lipka et al. [20] used machine learning techniques to identify featured articles using character trigram and part-of-speech trigram vectors. These features that are known to be characteristic of writing style out-performed alternatives in both a single domain and a domain transfer situation with F-measure scores of 0.88 across domains and good performance on articles of varying length.

Dalip et al. [8] presented a comprehensive assessment of quality indicators in collaborative content curation with a focus on Wikipedia. In their analysis they considered 69 indicators including text features, review features and basic network features. They used a machine learning approach to discover the most effective indicators and combination of indicators. They found that the easy-to-extract text-based features were most informative – more informative that more complex features based on link analysis.

This work by Dalip et al. [8] and Lipka et al. [20] is complementary to ours in that our network-based features can be combined with their content-based features to further improve classification accuracy.

3 Temporal Network View

In the network perspective proposed by Jurgens and Lu [13] a Wikipedia article is viewed as a sequence of revisions ordered by the edit times. These revisions can be tagged by edit types, such as *add* or *revert*. The source for this annotation can be seen in Figure 1. This shows a portion of the edit history of the Wikipedia page on "African Nationalism". It shows the date and time of the eight most recent edits, the resulting size of the page in bytes and the number of bytes added or deleted. According to the number of bytes changed in the edit and the comment provided by the editor, revisions are divided into five types: *major add*, *minor add*, *major delete*, *minor delete* and *revert*. An *add*-revision means the total number of characters in an article is increased after the revision is made. Conversely if the size of the article is reduced the revision is considered a *delete*-revision. A revision also can be seen as major or minor according to the absolute value of bytes changed. A major edit happens when a section of text has been inserted or removed, and a minor edit might be word spelling correction, adding or deleting few words. Jurgens and Lu [13] have shown that the correlation between edit size and page size is poor. Because of this they use a fixed threshold to separate minor and major edits rather than have a threshold that varies with page size. In our experiments we set 100 bytes as the minor/major threshold.

Some revisions are committed in order to revert an article to a previous stage, and in this case editors usually write comments to declare that the revision is a "revert". We mark an revision as a *revert* if the comment contains strings such as "revert", "rv" or "rvv". Clearly reverts are significant events as they indicate that all is not going smoothly in the article development. In the extreme case they can indicate an editing "war".

Jurgens and Lu [13] consider a further edit type which they simply call *edit*. This covers situations where an edit session results in the insertion *and* deletion of a number of tokens (words). We do not include this category because these edit events cannot be identified without fetching the original papers to compare tokens. The five types we consider can be identified simply be analysing the edit log reflected in Figure 1. Jurgens and Lu's *edit* events will typically be modelled as *minor add* or *minor delete* depending on the net impact on the page size.

Figure 1: A portion of the edit history for the Wikipedia page on "African nationalism". At the time of this snapshot the most recent edit entailed the addition of 155 bytes - this would be a *major add* in our notation.

3.1 Temporal Network Motifs

In this scenario an article is represented as a long sequence of edit symbols as shown in Figure 2 where there are five symbols in the vocabulary. A sequence of three *major add* symbols might represent healthy progress whereas a sequence of three *revert* events represents conflict between editors.

In contrast with the other network views, the motifs that best characterize this temporal data are sequence rather than network motifs. We consider motifs of length three with each motif covering a sequence of three revisions so a set of 125 temporal patterns is generated from the five edit types. In contrast with the work of Jurgens et al. [13] we did not capture whether the sequence revisions were edited by the same author as this results in more motifs. The temporal motifs are counted through the full revision history of an article with the total count of motif instances being equal to the total number of revisions minus two. The article is then represented as a vector of 125 motif counts representing the temporal motif distribution.

4 Ego Networks

Ego networks are static bipartite networks representing major edit events between authors and the ego node (the article) and also edits to other articles by those authors as shown in Figure 3 [26, 27]. To build these edit networks we applied several rules to filter the raw data. Firstly, as some articles have long edit histories, we only consider the latest 2000 revisions (most articles have less than 2000 revisions). The revisions are sorted by size and only the top 50% of revisions are considered further. The editors who made these major revisions are selected as "major editors" to include in the ego network. We also considered all articles that are connected by hyperlinks from the originating or ego-article and involved the editors who made the major contributions on these articles. We retained the linked articles that have been edited by at least one of the ego-article's major editors. Editors

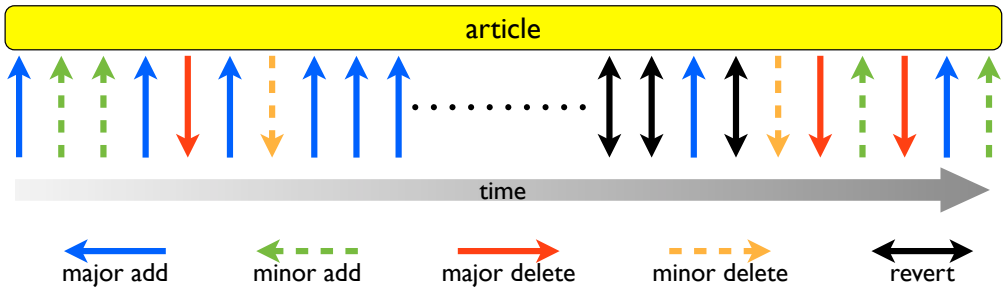

Figure 2: A temporal sequence showing the revisions of a Wikipedia article, illustrating five types of revision.

often repeatedly save their changes during a short session, so we judged continuous revisions by the same editor as a single revision.

In Wikipedia there are two type of editors, registered and unregistered users, where unregistered users are automatically labeled by the IP address they used when editing. We include both the registered and unregistered users to create the ego network.

Bots are allowed by Wikipedia to do some automatic editing and conventionally use names starting or ending with 'bot'. Bots perform a huge amount of small editing tasks so the bots are often very high-degree nodes in the networks. For this reason we dropped the bots from the networks as their high-degree nodes distort our network motif counting results. Furthermore we do not expect that the amount of attention from bots should impact on the quality of the article.

As with the temporal network representation, revisions can be *add* or *delete* actions. We represent an *add* action using an arrow from an editor to an article and an edge from an article to an editor means a *delete* revision – see Figure 3.

The ego network formulation only allows single edges between article and editor nodes so the direction of the edge is determined to be an *add* or *delete* based on the aggregate contribution of the editor to the article. So a directed edit describes the overall behaviour of an editor on an article.

4.1 Ego Network Motifs

In earlier work [27] we found that four basic undirected motifs, shown at the top in Figure 4, capture most of the classification power of five-node motifs. So here we extend these motifs to consider all variants with directed edges. Each undirected motif can produce several directed motifs representing the *add* and *delete* actions between articles and editors – in all there are 19 directed motifs extended from the 4 undirected motifs as shown in Figure 4.

In addition to these 19 motifs we also added three multi-edge motifs, which are shown in the right hand column in Figure 4. The objective with these additional motifs is to capture sequential behaviour of editors, i.e. pick up on sequences of major additions or add/delete cycles.

5 Trajectory Networks

Keegan et al. [15] present the idea of a trajectory network that is similar in some respects to the temporal network representation presented in Section 3. From the first revision to create a new article to the latest revision, these revisions can be linked as a path or trajectory. In a trajectory network, the nodes represent editors who made the revisions and a directed edge linking from editor A to editor B means that B's revision follows A's revision. Figure 5 demonstrates the construction of a trajectory network step by step. Editor A creates a new article following by editor B, so we link them from A to B. Editors may make contributions with a number of revisions and this is reflected in the *in-degree* count for that editor node. Self-looping edges are ignored so multiple consecutive edits by the same editor are not captured in this representation.

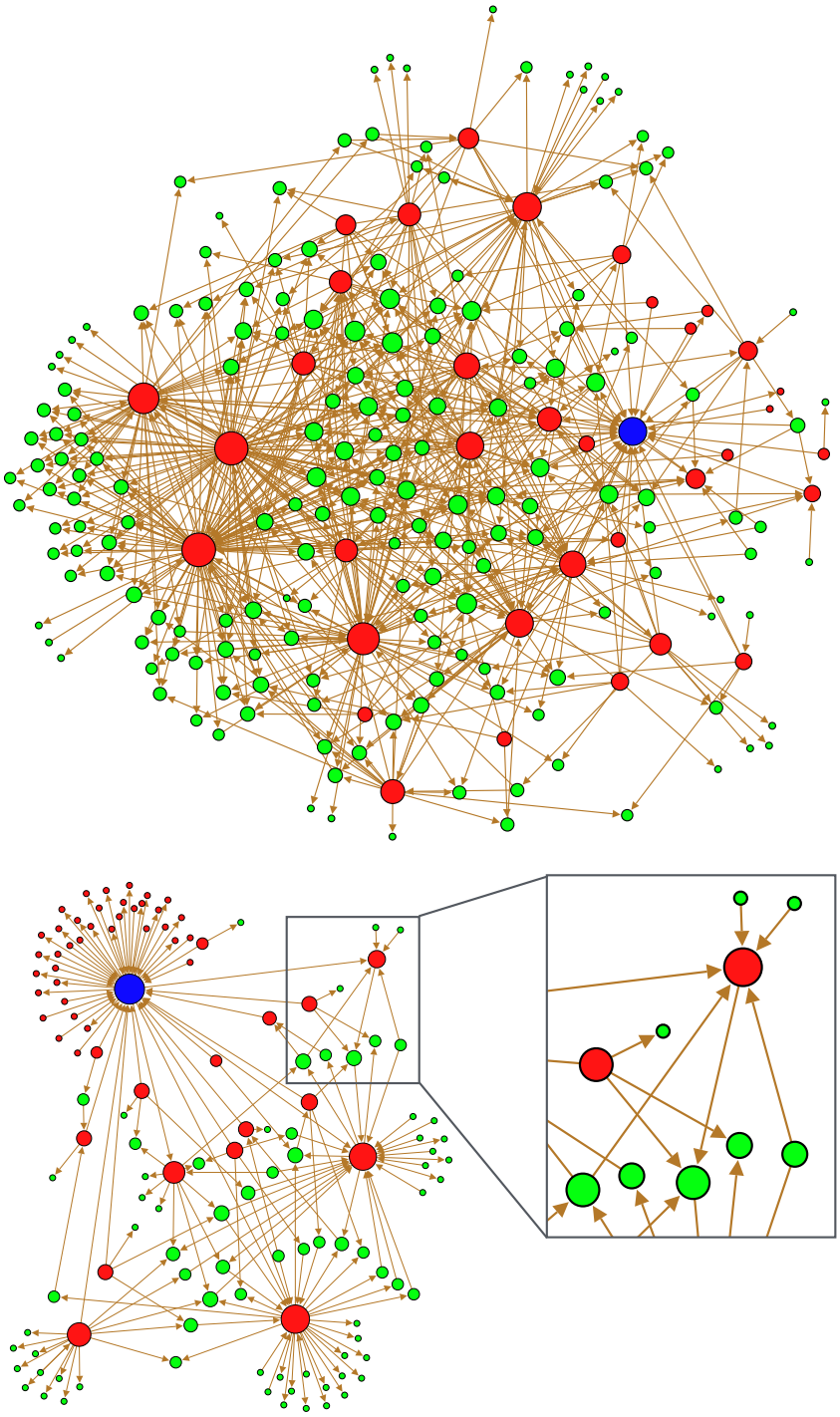

Figure 3: Two sample ego networks for "Aldfrith of Northumbria" (*Featured* article) on the top and "African nationalism" (*Start*-class article) on the bottom. The blue nodes are the ego-articles, and the linked articles are in green while editor nodes are shown in red. An out-edge from a red node (editor) is an *add*-action and an in-edge means a *delete*-action.

Trajectory networks do not always look like chains but can get quite dense when editors work on articles through a number of revisions at different times. Figure 6 shows trajectory networks for

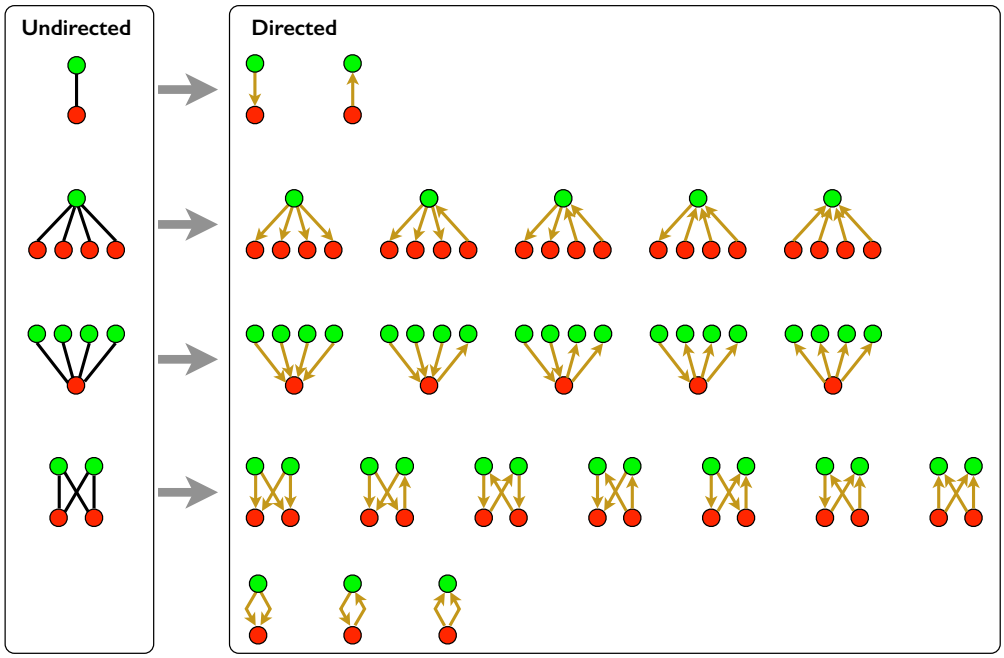

Figure 4: The 22 ego network motifs are used to analyse the ego networks, where 19 motifs are extended from the four undirected motifs we found in early work while 3 multi-edge motifs are also included to observe the behaviour of particular editors.

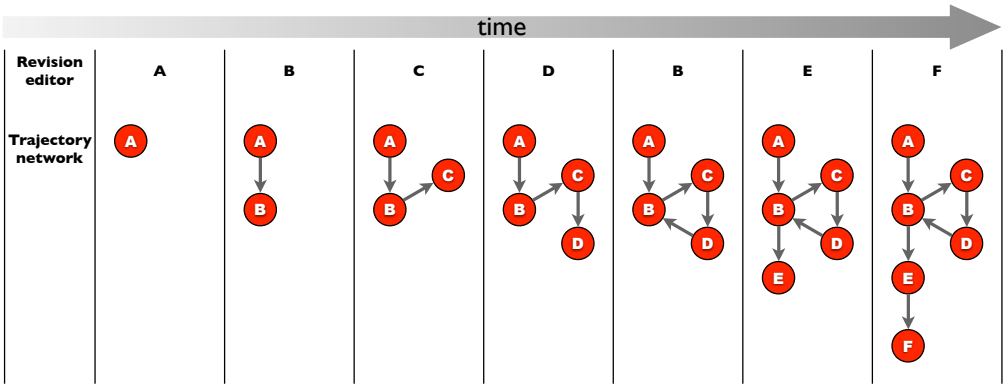

Figure 5: A simple trajectory construction example with 7 revisions made by 6 editors.

articles at two different quality levels, the network on the top represents a *Featured* article while the one on the bottom is a *Start* class article.

5.1 Trajectory Network Motifs

It is clear from a cursory analysis of the networks in Figure 6 that they differ in terms of some key statistics such as network diameter, clustering coefficient and betweenness centrality. While statistics such as these are the measures used by Keegan et al. [15] we continue with a motif-based strategy and use the triad network motifs introduced by Holland and Leinhardt [11] (see Figure 7). The 13 motifs cover all cases of three nodes connected with uni- or bi-directed edges. We did not extend to motifs of 4 or more nodes to keep the number of motifs reasonable as 4 nodes can produce 199 connected motifs.

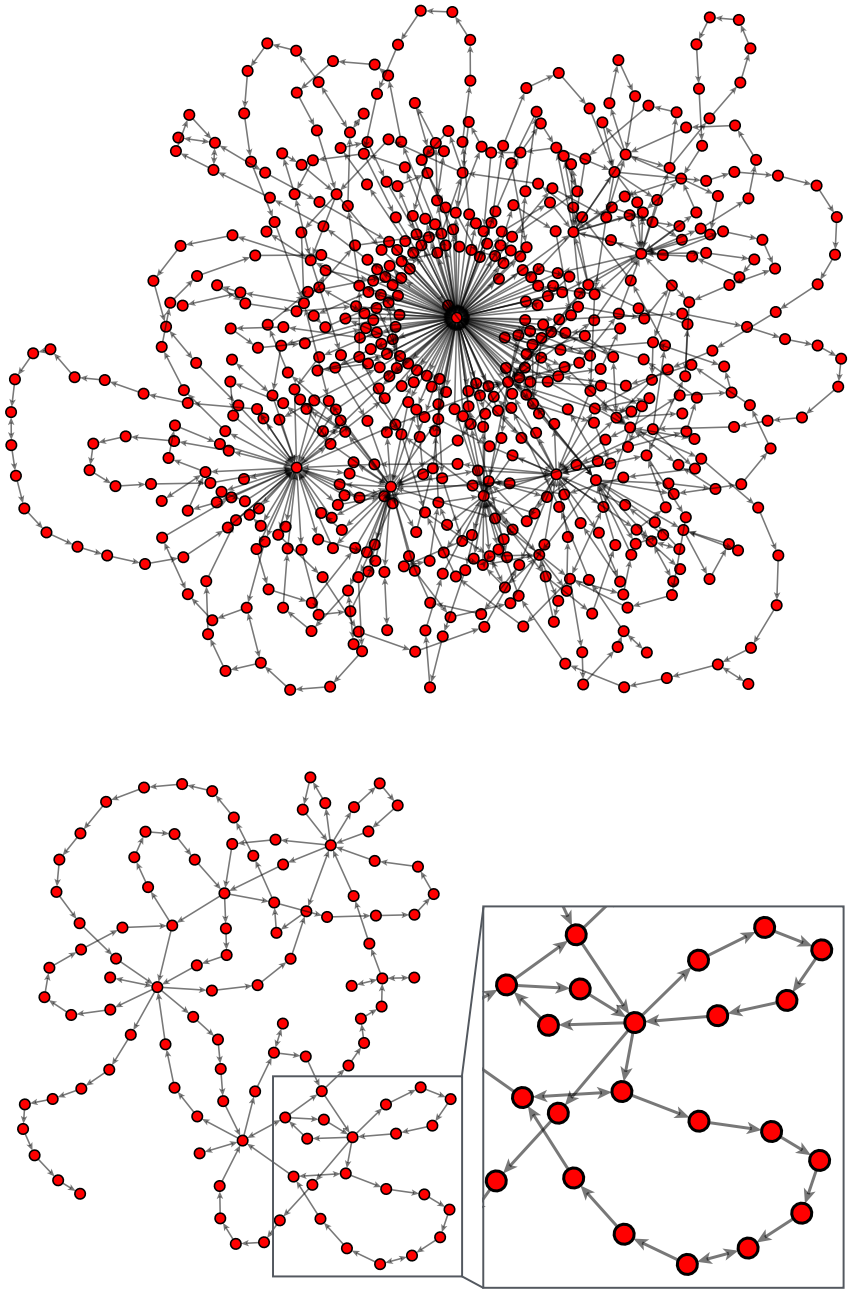

Figure 6: Two sample trajectory networks from Wikipedia. Nodes are the editors and the edges indicate edit sequence. The network on the top is a *Featured* article entitled "Chariot racing", and the one on the bottom is "African nationalism", a *Start*-class article.

6 Evaluating Relative Performance

Each of the three views provides a characterisation of a Wikipedia article as a vector of motif counts. The objective now is to assess the effectiveness of these motif profiles in predicting article quality. Each motif profile is a feature-vector representation that can be tested in a supervised learning framework to assess how predictive the different motif views are of the quality class. In all the tests that follow the classification is done using either random forest or logistic regression – in earlier work we found these to be the most effective classifiers with the ego-network representation [27].

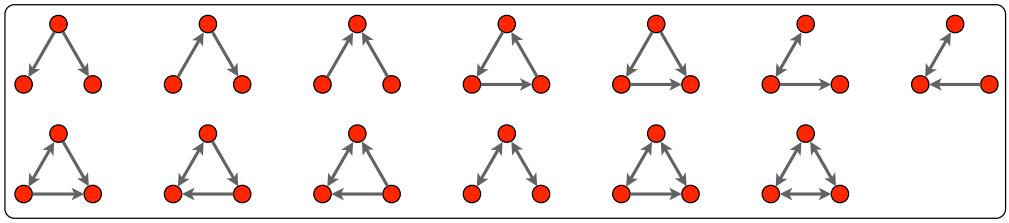

Figure 7: The 13 triad network motifs used to characterise the trajectory networks.

The test is to score Wikipedia article quality as defined by the Wikipedia quality scale[1] which runs from high to low as *Featured*, *A*, *Good*, *B*, *C*, *Start* and *Stub* (see Table 1). It is worth noting that the vast majority of Wikipedia articles are *Start* or *Stub* class. While we take *Start* articles as the core of our 'low' quality article sets, *Start* class articles account for a lot of Wikipedia use.

Table 1: Wikipedia quality classes. (Totals as of December 2013)

Class	**Total**	**Summary**
Featured	4,759	"Professional, outstanding, and thorough; a definitive source for encyclopedic information."
A	1,319	"A fairly complete treatment of the subject."
Good	20,127	"Useful to nearly all readers, with no obvious problems..."
B	91,910	"The article is mostly complete and without major issues..."
C	161,932	"The article is substantial, but is still missing important content or contains a lot of irrelevant material."
Start	1,066,789	"An article that is developing, but which is quite incomplete and may require further reliable sources."
Stub	2,387,604	"A very basic description of the topic."

We construct two classification tasks, an easy task that compares *Featured* articles with *Start* class articles and a harder task that compares *Featured* and *Good* articles against *C* and *Start* class articles. (The *A* class was skipped because there are few articles in that class.) We gathered Wikipedia articles from three large projects - *History*, *USA* and *Meteorlogy* in order to have enough *Featured* articles. In addition, we merged the articles from these three collections to build cross-domain datasets for both the easy and hard cases. The numbers of articles for each class and collection are listed below in Table 2.

Table 2: Datasets used in out experiments.

	Easy case		**Hard case**	
	Featured	**Start**	**FG**	**CS**
History	149	299	440	588
USA	272	299	564	597
Meteorology	131	298	431	598
All	552	896	1435	1783

6.1 Classifications using Individual Networks

Table 3 and Table 4 show the classification accuracies and ROC area on each of the eight classification tasks. The classification is performed using a random forest with 100 trees and accuracy is measured using 10-fold cross validation.

[1]Details on the Wikipedia Quality Scale are available at: `http://bit.ly/1avQfU`.

Table 3: Classification accuracy results using three different network views on both the easy and hard cases.

	Easy case				Hard case			
	History	**USA**	**Meteo**	**All**	**History**	**USA**	**Meteo**	**All**
Temporal	**92.0%**	89.1%	82.1%	87.8%	75.5%	**78.6%**	65.5%	**72.6%**
Ego	85.9%	86.5%	**83.0%**	84.2%	72.1%	74.5%	**70.5%**	69.9%
Trajectory	91.7%	**89.3%**	78.3%	**88.3%**	**75.8%**	74.1%	68.6%	72.2%

Table 4: Classification ROC area results using three different network views on both the easy and hard cases.

	Easy case				Hard case			
	History	**USA**	**Meteo**	**All**	**History**	**USA**	**Meteo**	**All**
Temporal	**0.976**	0.960	0.875	0.945	0.825	**0.853**	0.690	0.792
Ego	0.931	0.923	**0.901**	0.919	0.785	0.813	**0.768**	0.769
Trajectory	0.971	**0.966**	0.857	**0.950**	**0.833**	0.829	0.725	**0.805**

There are some clear patterns in the results that can be summarised as follows:

- The easy case is indeed easier than the hard case with classification accuracies higher by about 15% on average.That is reasonable as adding *Good* and *C*-class articles makes the task more difficult.
- Temporal and trajectory methods performed similarly with high accuracies and both of them beat the ego network method in almost all the scenarios.
- The Meteorology data seems to be harder to classify than the other datasets. One explanation for this may be lengths of the revision histories. Table 5 shows the average lengths of the revision histories for *Featured* and *Start* articles in each category. It seems that Meteorology articles reach *Featured* article status quickly compared to other projects.
- There is a clear correlation between simple accuracy and ROC area so accuracy only is considered in the remaining evaluations.

Table 5: The average number of revisions for *Featured* articles and *Start*-class articles, and the ratios between them.

	Featured	**Start**	**Ratio**
History	958	137	7.01
USA	1866	163	11.43
Meteorology	716	187	**3.82**
All datasets	1347	162	8.30

6.2 Comparison Against Content Features

Given that methods from text/document classification can be considered for this quality assessment task it is worth comparing the performance of our network-based features with alternative content based features. Dalip et al. [8] present a comprehensive evaluation where they consider three different types of features

- **Text features:** length, structure, style, readability.
- **Review features:** derived from editor and reviewer activity on the article.
- **Network features:** centrality, assortativity, clustering coefficient.

It is fair to say that these review features set out to capture similar information to the temporal network features in our study; however the temporal network characterisation should be more comprehensive. Similarly, our ego-network features should cover the collection of network features in their study.

They present results on a regression study where the quality classes map to numeric values as follows: {Featured Article, A Class, Good, B Class, Start, Stub} map to {5, 4, 3, 2, 1, 0}. They do not consider C class articles as that category did not exist when they gathered their data. On this regression task, they report mean squared error (MSE) scores between 0.92 and 1.19 for different combinations of their text features.

With our data we construct a similar regression task with the following mapping {Featured Article, Good, C Class, Start} $\rightarrow$ { 4, 3, 2, 1} on all the articles listed in Table 2. A ten fold cross validation on this regression task returns an average MSE is 0.96. This is in the same range as the performance of the text features in the study by Dalip et al.

7 Integration Strategies

One thing that emerges from the descriptions of the different views in Sections 3, 4 and 5 is that there are aspects of behaviour captured by some views that are not captured by others. Thus, if we integrate the three views it should result in improved performance on our Wikipedia quality classification task.

This fusion of different perspectives has received a fair deal of attention in the research literature and there are two clear alternatives [22][10]. The integration can be at the *feature* level which is called early integration or at the *decision* level which is termed late integration [10]. Early integration simply entails concatenating the feature vectors from each view into a composite vector and using that as input to a classifier. Late integration entails training classifiers for the individual views and then integrating the output of these classifiers, typically in an ensemble framework.

In order to assess the effectiveness of these integration strategies we conduct an evaluation using just two views, the temporal view and the ego network view. Early integration results in a vector of 147 features produced by combining the 125 temporal motifs with the 22 ego network motifs. This representation is then tested with logistic regression and random forest classifiers.

Late integration is achieved using an ensemble of 20 members, 10 trained using temporal motifs and 10 trained using ego network motifs. In order to ensure diversity in the ensemble the classifiers are trained using feature subsets selected at random so that each subset contains two thirds of the features. The decisions of the individual classifiers are integrated through majority voting with the majority class (*Start* or *CS*) selected in the event of a tie.

7.1 Performance of Early & Late Integration

The experiments for both strategies were performed using random forest (100 trees) and logistic regression classifiers on the combined datasets (*All* in Table 2). Results for classification accuracy measured using 10-fold cross validation are shown in Figure 8.

The results are consistent across both datasets. For each classifier type, early integration is better than late integration. Early integration using random forest is the best and late integration using logistic regression performs worst. So in the experiments that follow we continue with early integration using random forest.

8 Integration of Three Views

In this section we present Venn diagrams to demonstrate the improvements in accuracy from integrating two and three network views. The Venn diagrams in Figure 9 show the performance on the integrated datasets. Appendix A shows the results for the remaining datasets.

The Venn diagrams are straightforward to interpret. The figures for classification using the individual views only correspond to those in Table 3. Performance using all three views is 91% in the easy case and 78.7% in the hard case. The improvement in classification accuracy in the hard case is impressive representing an increase of 5.5% over the best single view.

An examination of all eight Venn diagrams shows that all three views make a contribution with no view being redundant. This is evident because in all cases, there is no combination of two views that

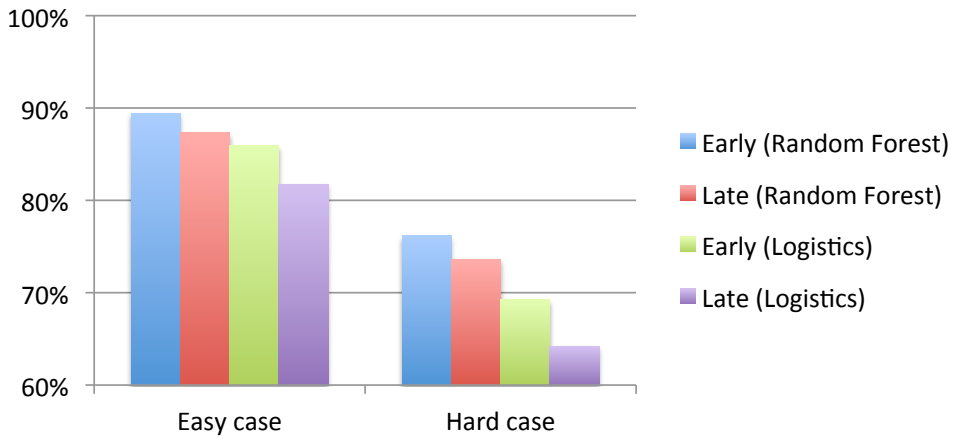

Figure 8: Classification performance of the two alternative integration strategies using logistic regression and random forest as the classifier on the integrated datasets.

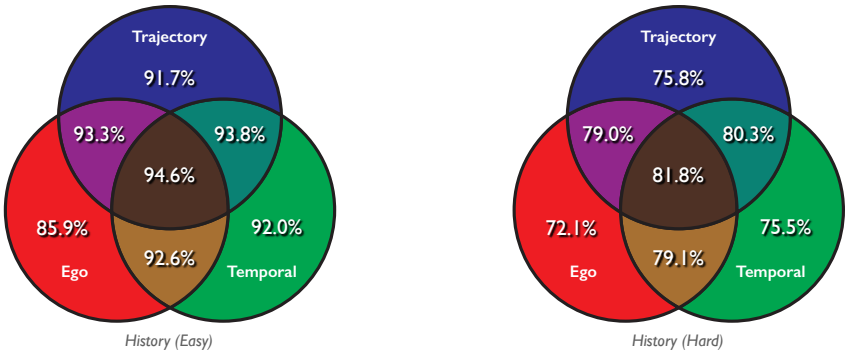

Figure 9: Classification performance using one, two and three network views on the mixed datasets.

is as effective as all three. In fact the unique contribution of an individual view can be quantified in this way as illustrated in the radar charts in Figure 10.

For instance, in the easy case (the left Venn diagram) in Figure 9, the classification accuracy using the integration of ego and temporal network views is 89.2% and the full-view integration achieved 91%, so the unique contribution of the trajectory network view is 1.8%. This point is plotted on the radar chart in Figure 10. These radar charts offer a number of insights into the integration process. For instance the improvements due to integration are greatest on the harder classification task. Also, the ego network makes its biggest contribution on the Meteorology data.

8.1 Feature Importance

The radial plots in Figure 10 show that all network views contribute to classification performance. We can analyse this further by scoring the predictiveness of individual features using information gain [28, 29]. Figure 11 shows histograms of the 40 top scoring features on the integrated classification tasks.

The most obvious point to note is that the features from the ego networks are not contributing much on the Easy task but are more prominent for the Hard task. On the Easy task the highest scoring features are triads 11, 7 and 6 from Figure 7. These are all open triads with bi-directional edges.

The same three motifs take second, third and fourth place on the Hard classification task with a temporal network motif taking first place. This winning motif is the sequence of three major add

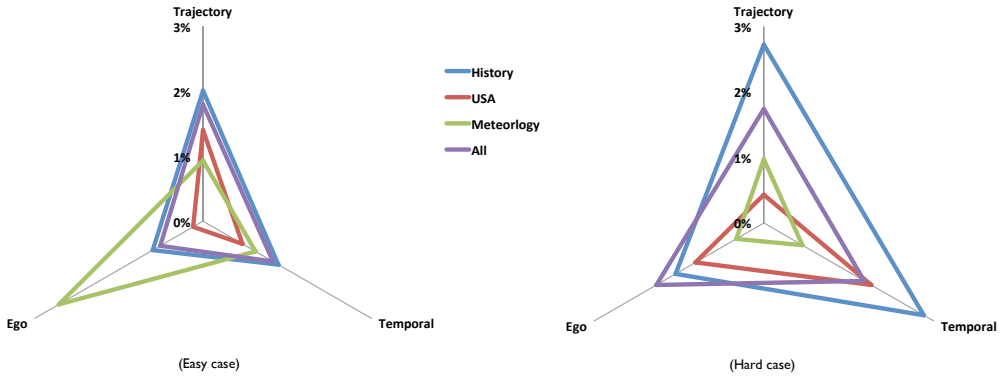

Figure 10: These radial plots quantify the unique contribution of each view to the integrated performance. The radius for a view quantifies the drop in accuracy when that view is not included – for instance there is a 1.8% drop in accuracy in the Easy *All* dataset when the Trajectory view is not included.

edits in a row. The next two motifs in the ranking are Ego motifs, the first two multi-edge motifs shown at the bottom of Figure 4.

9 Summary and Conclusions

Determining appropriate network representations for dynamic collaborations such as the development of Wikipedia pages is a significant challenge. In this paper we have described three alternative network representations for collaboration among Wikipedia editors and proposed a quality classification task in order to assess the effectiveness of the representations. It is clear from the evaluation that the different network views capture different aspects of the interaction so we have evaluated alternative strategies for integrating the views in classification. The evaluation shows that early integration using random forest was the best solution for the classification task.

In future work we will seek to develop an integrated representation that captures representation power of the three strategies. We will perform a feature selection analysis across the integrated representation to identify which motifs (associated behaviors) are providing the predictive power.

Acknowledgments

This work is supported by Science Foundation Ireland Grant No. 08/SRC/I140 (Clique: Graph & Network Analysis Cluster).

References

[1] B. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, page 270. ACM, 2007.

[2] B. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the Wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, pages 1–10. ACM, 2008.

[3] E. G. Allan, Jr., W. H. Turkett, Jr., and E. W. Fulp. Using network motifs to identify application protocols. In *Proceedings of the 28th IEEE Conference on Global Telecommunications*, GLOBECOM'09, pages 4266–4272, Piscataway, NJ, USA, 2009. IEEE Press.

[4] R. Baeza-Yates. User Generated Content: How Good Is It? In *3rd Workshop on Information Credibility on the Web (WICOW 2009)*, pages 1–2, 2009.

[5] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 16–24, New York, NY, USA, 2008. ACM.

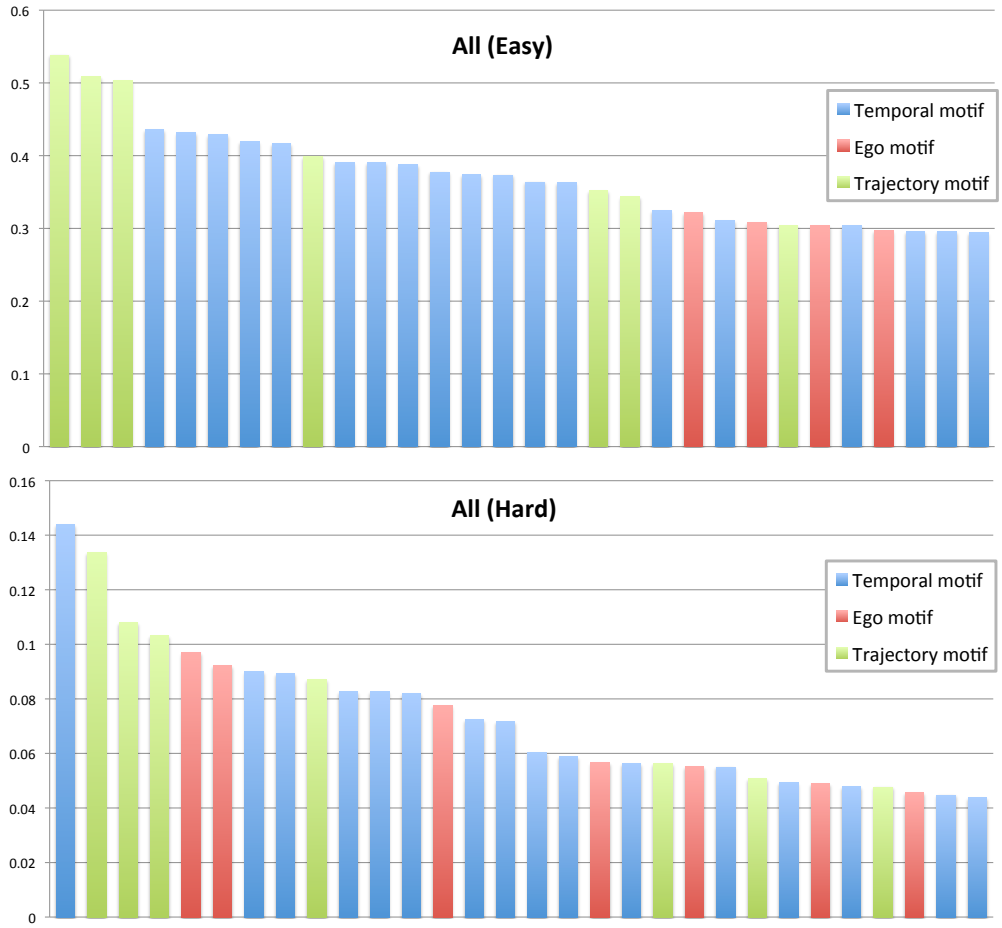

Figure 11: Histograms of the information gain scores for the top 40 features in the *All* dataset for the Easy and Hard tasks. The Ego motifs don't score well for the Easy task but figure more strongly in the Hard task.

[6] P. Boykin and V. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61 – 68, 2005.

[7] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 731–740. ACM, 2009.

[8] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 295–304. ACM, 2009.

[9] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[10] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pages 3437–3443. IEEE, 2005.

[11] P. Holland and S. Leinhardt. Local structure in social networks. *Sociological methodology*, 7(1), 1976.

[12] T. Iba, K. Nemoto, B. Peters, and P. Gloor. Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6441–6456, 2010.

[13] D. Jurgens and T.-C. Lu. Temporal motifs reveal the dynamics of editor interactions in wikipedia. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.

[14] K. Juszczyszyn, P. Kazienko, and K. Musiał. Local topology of social network based on motif analysis. In I. Lovrek, R. Howlett, and L. Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5178 of *Lecture Notes in Computer Science*, pages 97–105. Springer Berlin / Heidelberg, 2008.

[15] B. Keegan, D. Gergle, and N. Contractor. Staying in the loop: Structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration*, 2012.

[16] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[17] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[18] D. Laniado and R. Tasso. Co-authorship 2.0: Patterns of collaboration in Wikipedia. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 201–210. ACM, 2011.

[19] A. Lih. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *In Proceedings of the 5th International Symposium on Online Journalism*, pages 16–17, 2004.

[20] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1147–1148. ACM, 2010.

[21] A. Narasimhamurthy, D. Greene, N. Hurley, and P. Cunningham. Partitioning large networks without breaking communities. *Knowledge and information systems*, 25(2):345–369, 2010.

[22] W. Noble et al. Support vector machine applications in computational biology. *Kernel methods in computational biology*, pages 71–92, 2004.

[23] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[24] F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system. In *PASSAT 2011, Third IEEE International Conference on Information Privacy, Security, Risk and Trust*, pages 1318–1326. IEEE, 2011.

[25] J. Surowiecki, M. Silverman, et al. The wisdom of crowds. *American Journal of Physics*, 75:190, 2007.

[26] G. Wu, M. Harrigan, and P. Cunningham. Characterizing Wikipedia pages using edit network motif profiles. In *(SMUC 2011) 3rd International Workshop on Search and Mining User-Generated Contents*, pages 45–52, New York, NY, USA, 2011. ACM.

[27] G. Wu, M. Harrigan, and P. Cunningham. Classifying Wikipedia Articles Using Network Motif Counts and Ratios. In *(WikiSym 2012) 8th International Symposium on Wikis and Open Collaboration*, 2012.

[28] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[29] R. Zhang and T. Tran. An information gain-based approach for recommending useful product reviews. *Knowledge and Information Systems*, 26(3):419–434, 2011.

A Venn Diagrams

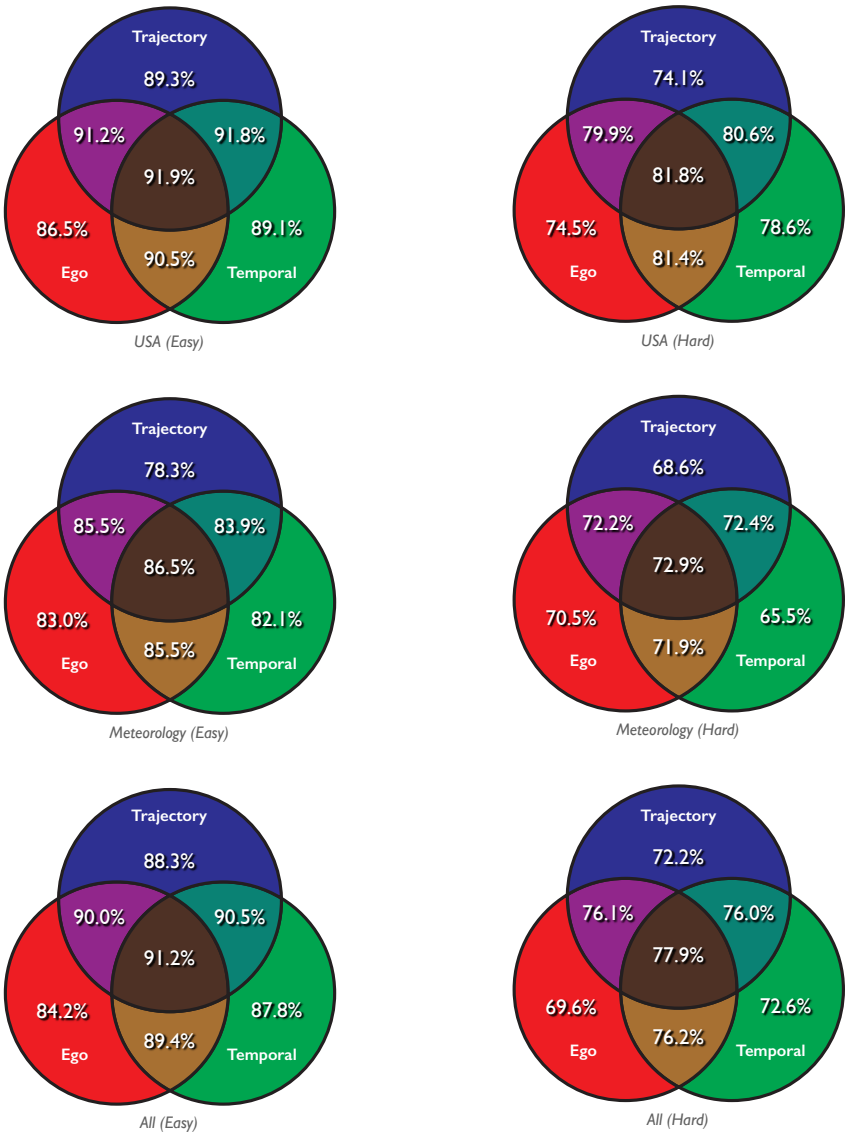

Figure 12: Venn diagrams for the individual datasets (*History*, *USA* and *Meteorology*) in both the easy and hard cases illustrate the individual performance and integrated performance for the three network views.