whoVIS: Visualizing Editor Interactions and Dynamics in Collaborative Writing Over Time

Fabian Flöck
Computational Social Science Group
GESIS - Leibniz Institute for the Social Sciences
fabian.floeck@gesis.org

Maribel Acosta
Institute AIFB
Karlsruhe Institute of Technology
maribel.acosta@kit.edu

ABSTRACT
The visualization of editor interaction dynamics and provenance of content in revisioned, collaboratively written documents has the potential to allow for more transparency and intuitive understanding of the intricate mechanisms inherent to collective content production. Although approaches exist to build editor interactions from individual word changes in Wikipedia articles, they do not allow to inquire into individual interactions, and have yet to be implemented as usable end-user tools. We thus present whoVIS, a web tool to mine and visualize editor interactions in Wikipedia over time. whoVIS integrates novel features with existing methods, tailoring them to the use case of understanding intra-article disagreement between editors. Using real Wikipedia examples, our system demonstrates the combination of various visualization techniques to identify different social dynamics and explore the evolution of an article that would be particularly hard for end-users to investigate otherwise.

Categories and Subject Descriptors
[Collaborative and social computing]: Collaborative content creation; [Human-centered computing]: Visual analytics

Keywords
Visualization, Interface, Graph models, Social Networks, Wikipedia, Online Collaboration, Social Dynamics

1. INTRODUCTION
Wikipedia as a socio-technical system has received its fair share of attention over the last decade. Yet, understanding the collaborative writing history of an article as a casual user, editor or even researcher in an easy, intuitive way (i.e., without relying on elaborate statistical analysis) is still a hard task. There is a lack of transparency regarding the editing process on Wikipedia: it is fully documented in the revision history, but not in a way that is straightforward to browse, inspect and analyze by humans in all its intricacy. For instance, one cannot easily discover which words were contributed by what author or what specific dynamics governed the rise of disagreement between editors on particular content in the article. This information would be key to enable accountability and social transparency, as has been argued by Suh et al. [4], but is hidden from the user due to the innate complexity. Some related visual interfaces as, e.g., “Wikidashboard”, “Wikitrust” and community solutions have been proposed.¹ Still, tools that allow users to visually explore the dynamic relationships between editors that emerge from the main activity in the system – the collaborative process of adding, deleting and restoring specific content – are not available or not equipped for the purpose of accurately reflecting all relevant interactions of editors with each other and the content.

In this work, we therefore construct and visualize editor-editor networks over time (per revision), derived from the collaborative editing actions on the word-level of single articles.² We introduce whoVIS, a novel, interactive Web tool for investigating the collaborative writing process of a Wikipedia article that combines all of the following features: (i) mining (re)introduction and delete actions of editors on each other’s written text at word granularity with proven accuracy, to infer and model editor-editor disagreement, (ii) a custom graph-drawing method for disagreement edges in Wikipedia editing that offers a meaningful depiction of the ongoing disputes in the article, (iii) an interface for interactively exploring the emerging network graph in a specific article over revisions, enriched with meta-information on editors and edges, (iv) a drill-down feature to learn how a specific edge between editors was constructed and which words were disagreed about (“edge context”) and (v) several auxiliary metrics over time that can be employed for better exploration and understanding of the main network graph.

The principal contribution of this work is a working and usable system built on top of established techniques to mine and visualize Wikipedia interactions, which are enriched with new features tailored for this use case (in particular, revision-wise graph exploration over time, and edge context). In doing so, we showcase how authorship and interaction mining from revisioned, collaborative writing can be transformed into a useful visual interface for exploring social dynamics and the provenance of content.

A demonstration of the system is available online at http://km.aifb.kit.edu/sites/whovis/.

2. RELATED WORK
Some scientific works treat the construction of intra-article, editor-to-editor networks based on the edit actions of users. Suh et al. [5] construct and draw networks of editors where a “negative” edge (u,v) indicates editor u completely reverting a revision submitted

¹(a) For community tools see http://en.wikipedia.org/wiki/Wikipedia:Tools. (b) Wikidashboard [4] visualizes edits over time by contributors, but does not track changed content. (c) Wikitrust provides word provenance information, but not interactions of editors.http://wikitrust.soee.ucsc.edu/.
²In contrast to inferring editor networks based on article co-editing.
by v, removing v’s content. Still, in [5] disagreement is detected only where an editor resets the article content to an exact duplicate of an earlier revision. Although this method covers a large part of edit actions, it does not comprise all reverts and disagreement actions (e.g., partial reverts). Consequently, Brandes et al., who initially proposed a similar approach, recognize in later work [1] that this method does “not consider who deletes how much of whose edits or who restores whose edits deleted by whom. However, [...] it is exactly this information that enables us to characterize individual authors and groups of authors”. To enable a more fine grained network construction and depiction, they hence improve this method in [1] to infer an edge (v,u), weighted with the exact words written by editor u that were subsequently (dis)agreed on by editor v.

Similarly, Maniu et al. [3] infer a signed network of positive (agree) and negative (disagree) relations between editors by (among other relationships) extracting changed words via text deltas. However, to verify if an edit actually constitutes a revert to a former other relationships) extracting changed words via text deltas. However, to verify if an edit actually constitutes a revert to a former

None of these approaches have seen an implementation as an interactive web interface for revision-wise network exploration yet, and neither offers possibilities for exploring from what content dissent specific edges were constructed.

3. NETWORK CONSTRUCTION AND VISUALIZATIONS

3.1 Mining of editor-text interactions

As a basis for the construction of the editor interaction network, we use the “WikiWho” algorithm [2] that computes, per revision, the authorship of each word (token) within a Wikipedia article. WikiWho has been shown to fulfill this task more accurately (avg. 95% correct authorship attribution) and efficiently than previous solutions. It thereby can guarantee a certain degree of quality of the base data we build our editor interaction extraction upon. In this work, we extend WikiWho to track which editors add, delete and reintroduce which tokens of the article text. In case of deletion, our extension of WikiWho monitors who has originally, i.e., for the first time, written the text; in case of reintroduction, it keeps track of the original author and the deleter of the restored text, as well as any additional delete or reintroduce actions by any later editors.

3.2 Network model - Constructing interaction edges between editors over time

To model how the edit actions of an editor relate her to other editors, we extend the model introduced by Brandes et al. [1]. Yet, as outlined in Section 1, we aim to provide the editor interaction network graph for each revision of the article to illustrate its evolution over time, instead of showing only the state of the network at the last available revision. Moreover, even when drawing the network of every revision r_i in the sequence, merely aggregating for all editors all interaction edges stemming from all previous revisions r_1,...,r_(i-1) is not sufficient for our visualization use case for two reasons. (i) The network graph quickly grows to be cluttered with a high amount of nodes and edges (already at around 70-90 revisions for most articles) and it thus becomes impractical for a user to distinguish interactions between individual editors or see patterns in a sub-graph as nodes and edges highly overlap. (ii) The “current state” of recent interactions patterns for a revision is not (easily)

obervable if graph elements generated in older revisions never disappear and thus can be hardly distinguished from recently created elements that might highlight a currently more relevant editing dynamic. On these grounds, we draw the editor network for each revision r_i by excluding actions previous to a threshold (window) r_{i-w}; we used \omega = 50 for the presented implementation. In the following, we provide a formal definition of the network interactions employed in this work. Given a Wikipedia page p and its history of revisions r_1,...,r_N, the edit interaction network associated with p is defined as an N-tuple G_p = (G_1,...,G_N) where \omega is the window size and each G_i is defined as follows:

- G_i = (V_i, E_i, \alpha_i, w_i) is the graph of interactions occurring within the window, i.e., between revisions r_{i-w} and r_i.
- The nodes V_i correspond to editors that have done at least one edit on p between r_{i-w} and r_i, or authors with at least one word originally written by them still present in r_i.
- The set of edges E_i \subseteq V_i \times V_i encodes the edit interactions among editors. An edge (u,v) \in E_i for editor u performed one of the following actions towards v:
  (a) u deletes text that has been originally written by v; the number of words deleted by u at revision r_j (r_j < r_i) is denoted as delete(u,v);
  (b) u undoes a delete by v by reintroducing the deleted text; the number of words restored by u in r_i, deleted by v at revision r_j (r_j < r_i) is denoted by undo_delete(u,v);
  (c) as an extension to [1] we include a further crucial relation that often appears in conflicts or after vandalism: u undoes a reintroduction of text by v, while the text originally could have been written by a different author; undo_reintro(u,v) denotes the number of words deleted by u in revision r_i, reintroduced by v at revision r_j (r_j < r_i). This type of action is always linked to a delete action that triggers the creation of an edge (u,w) \in E_i, since u deletes words originally written by w during the undo of the reintroduction of v in r_i.
- \alpha_i : V_i \rightarrow \mathbb{N}_0, for each node u \in V_i, \alpha_i(u) corresponds to the number of words in r_i that are originally authored by u.
- w_i : E_i \rightarrow \mathbb{N} corresponds to the edge weight function. It measures the disagreement between editors u and v. For each (u,v) \in E_i, w_i(u,v) is calculated as follows:

\[ \sum_{j=r_i-w}^{r_i} \text{delete}_j(u,v) + \text{undo_delete}_j(u,v) + \text{undo_reintro}_j(u,v) \]

In accordance with Brandes et al. [1], these relationships are interpreted as disagreement (negative) edges between editors. Positive relationships are not included for this work. 

3.3 Auxiliary metrics

We define several metrics that can help to guide a user by (i) providing additional information about the editor relations and patterns explorable in the interaction graph (described in Section 3.5) to better understand their meaning and (ii) by highlighting potentially interesting phases in the development of the article for target-oriented navigating of the sequence of network states per revision. Given a graph G_i = (V_i, E_i, \alpha_i, w_i) of the edit interaction network, we define: (i) Number of Disagreement Actions: The total number of negative actions, computed as the sum of all the values in w_i. (ii) Bipolarity [1]: Degree of how well editors are divided into two poles of disagreement. (iii) Authorship Gini-Coefficient: 5

5Although “agreement” relations between u and v (e.g., “restore” and “redelete”) can be inferred by our method, we reserve this extension for future work, as they are neither crucial for the used drawing method (cf. Section 3.5), nor easily combinable with it. 
Measures how equal the authorship of tokens is distributed over the editors that have contributed to the content of revision $r_i$. Let $e$ be the sequence of the $n$ editors that own at least one token in $r_i$, indexed in non-decreasing order of authorship ($\alpha_i$), we define: 

$$\text{authorship_{gini}} = 2 \frac{\sum_{i=1}^{n} \alpha_i e_i \sum_{j<i} \alpha_j}{\sum_{i=1}^{n} \sum_{j<i} \alpha_i \alpha_j} - 2 \frac{1}{n}$$

(iv) Disagreement Focus: High values indicate that the negative actions performed in $r_i$ by $u$ are particularly targeting editor $v$; calculated as: 

$$\text{focus}_{u,v} = \frac{\sum_{w \in \text{tokens}} |u, w, v, w|}{\sum_{w \in \text{tokens}} |u, w, v, w| + |v, w, u, w|}$$

(v) Reciprocity: Mutual disagreement over time between editors $u$ and $v$ in $r_i$, denoted as reciprocity,$(u, v)$, is modeled as a weight function (weight $\phi \in [0.0; 1.0]$) of the portion of content in disagreement between $u$ and $v$ in $r_i$ and the average disagreement focus of $u$ and $v$ in the window $\omega$: 

$$\phi \cdot \frac{\text{min}(w_{(u,v)}, w_{(v,u)})}{\text{max}(w_{(u,v)}, w_{(v,u)})} + (1 - \phi) \cdot \text{avg}_{i \leq j \leq \omega}(\{\text{focus}_{u,v}, \text{focus}_{v,u}\})$$

### 3.4 Edge Context: Explaining Disagreement

In existing solutions for depiction of editor-interaction in Wikipedia it is close to impossible for a user to understand what exactly editors were (dis)agreeing about from the plain network edges and target revisions for the action are displayed, and a link is given to the Wikipedia “diff” for the revision.

#### 3.5 Visualization Implementation

The visualization is implemented using D3. After selecting an article, users visualize the editor network and a plot of the metrics from Section 3.3. Users can navigate the editor network over time via a slider or skip buttons. In addition, whoVis offers two more view tabs: “ownership”, and “additional metrics”. Given a revision $r_i$, the ratio of a node $u \in V_i$ in the graph is proportional to the percentage of words in $r_i$ that were authored by $u$, i.e., $\text{ratio}_{u}(u) = \alpha_i(u) / \sum_{v \in V_i} \alpha_i(v)$. A minimum ratio is defined to make nodes visible even if the editor did not author any text. Every node is assigned a color; this allows for easily tracking nodes when their position changes over time. The node of the editor of $r_i$ is highlighted with a dark-colored border. Nodes can be dragged to clear any potential overlap of graphical elements; hovering over a node will highlight all connected nodes.

The coordinates of nodes are computed with the approach by Brandes et al. [1]. This technique is the most fitting as nodes are unvaringly placed in the center of the graph if they are neutral to each other or only do small corrections, while high disagreement nodes get “pushed out” to the periphery. The algorithm computes the eigenbasis of the matrix $A$, $A(u, v)$ being the disagreement between editors $u$ and $v$, and vice versa. We build a matrix $A_r$ for each $r_i$, with $A_r(u, v) = w(u, v) + w(v, u)$, then, the two most negative eigenvalues of $A_r$ and their eigenvectors $x_i$ and $y_i$ are computed, resulting in the $x$- and $y$-coordinates of nodes in $r_i$, respectively. Editors in window $\omega$ that did not cause a disagreement...
edge (e.g., by just adding content) are not displayed in the main graph but on a separate "non-disagreeing editors" line, to keep track of recently active editors. An additional row lines up all nodes that represent authors of any content in \( r_i \) that did not edit in \( \omega \) at all.

An edge \((u, v) \in E_i\) is drawn as a grey line with width proportional to the value \( A_i(u, v)\). The edge coloring is changed to red if the disagreement is mutual, i.e., values \( w_i(u, v) \) and \( w_i(v, u) \) are both \( > 0 \). The opacity of the red color starts with minimal value and increases according to the reciprocity metric.

4. USAGE AND USE CASES

As an example, we look at the article Tropical Storm Alberto (2006), given in [1] as an instance of a "featured" (i.e., high quality) article with low bipolarity, meaning that the aggregated network of the last revision exhibits multipolar disagreement structures instead of, e.g., two dominant disagreeing groups (cf. Figure 2 in [1]). Exploring the history of interactions in whoVIS (Figure 2), we can first see in the line chart under the network graph that the bipolarity of the network can, in fact, be very high at times, even when the aggregated graph for the last revision shows low bipolarity. This can be explained by the fact that over time, we see ephemeral bipolar disagreement "camps" of different editor combinations emerge and disappear. For instance, the high bipolarity spikes at SrevID \( \approx 280 \) (SrevID = "sequential revision id", assigned by whoVIS for this article) and SrevID \( \approx 360 \) indicate disagreements between editors CrazyG83 and Super-Magician, while the spike at, e.g., SrevID \( \approx 520 \) shows a disagreement between Hurricanehink vs. Thegreatadr and Titoxd (see Figure 2). The revision-wise exploration hence allows us to deconstruct and better understand the aggregate disagreement network in terms of its dynamic evolution by identifying temporal sub-structures of disagreement.

Going through the network graph chronologically, we can see that after the foundation of content by CrazyG83, a phase of indicated disagreement between several editors follows starting at about SrevID \( \approx 80 \). We can see mutual disagreement mainly between CrazyG83 and Super-Magician, with several editors entering into a "disagreement triangle" with them before the dissent dies down towards SrevID \( \approx 280 \) (Figure 1 shows an intermediate step). We observe this development mirrored in the average reciprocity and total negative actions charts. Inquiring into the (mostly highly reciprocal) disagreement edges via edge context in this phase reveals that the mutual editing of the actors, for the most part, concerns updates relating to recent developments of the titular "Storm Alberto" rather than a major clash of subjective viewpoints. This can be gleaned from the actions performed (largely date-related updates), the comments ("7 p.m. update") and the high edit rapidness (via the corresponding line chart in "additional metrics"). Yet, mixed into this "live reporting" spurt are genuine opinion clashes about how to write the article, e.g., the disagreement edge emerging between Derek.cashman and Super-Magician at SrevID \( \approx 184 \), arguing whether to include links in section headers and citing the pertaining Wikipedia policies, as illustrated in Figure 1. Later, we see disputes about, e.g., the veracity of a report between Weatherreport111 vs. Ajm81 and Hurricanehink (SrevID \( \approx 470 \)); and vandalism fighting, as surfacing at SrevID \( \approx 648 \) between the IP 190.51.x.x and several registered users amidst other, content-related disagreements, which develop (quite literally and visually) orthogonal to the vandalism fight (cf. Grisou492 vs. Julianoffton around the same time).

These examples showcase, with the help of the edge context and revision-wise exploration, that "disagreement", modeled as text-deletes and -reintroductions can have highly different meanings in specific situations and in fact moves on a spectrum between mere "corrections", "profound disagreement" and "outright conflict", a distinction that can be easily overlooked when boiling down real human editor interactions into statistical graph representations. These different disagreement types can overlap, co-exist in parallel or appear at different points in time. Edge context hence enables a crucial qualitative assessment of editor interactions by augmenting the information captured in the network graph.

Another interesting observation in the article is the development towards "featured article" status, which it reaches at SrevID = 584 (captured in "additional metrics"). The Authorship Gini curve shows a significant increase in authorship concentration before that event, at SrevID \( \approx 510 \) (Figure 2), which, upon inspection of the "word ownership" of the top authors in the respective whoVIS tab, can be attributed to a large "writing sprint" by user Hurricanehink. This editor contributes a vast amount of content with some deletion/rewriting of authors Titoxd, Thegreatadr and Ajm81, but without being antagonized, and mostly adding new material of his own. After reaching featured article status, however, we see a burst of disagreement following SrevID \( \approx 635 \), which is caused by many new editors doing small corrections but also partly due to vandals appearing, e.g., at SrevID \( \approx 648 \) (IP 190.51.x.x.). By tracking authorship concentration, the top editors’ authored content and other metrics over time, as well as monitoring important wiki-templates in the article, whoVIS can inform a targeted inspection of editing interaction dynamics in the network by indicating significant phases in the article lifecycle through aggregate metrics over time.

5. CONCLUSIONS AND FUTURE WORK

In conclusion, we have indicated that a variety of disagreements patterns exists and that their evolution over time merits investigation. We further showed how the innovative data presentation of whoVIS can help discover those intricate details of collaborative writing that would else be hard or impossible to glean from existing end-user tools. We are not aware of a tool solution that provides the revision-wise exploration of the interaction network in conjunction with the qualitative exploration offered through edge context and auxiliary metrics (especially authorship tracking) that together allow for novel views of collaborative writing dynamics.

Our approach could also be adapted to non-Wiki environments as, e.g., code repositories like GitHub, which might similarly benefit from a visual analysis of their collaboration patterns. Other future developments of our tool will likely include (i) automatically and visually distinguishing types of disagreement, (ii) inclusion of positive edges describing support of another editor’s content, and (iii) a more comprehensive version of the edge context.

6. REFERENCES