Växjö universitet

Social Visual Analysis and Interaction

Seminar ”Visual Analytics” (DA 4024)

Tatiana Braescu
June 01, 2008
Växjö University, MSI
Advisor: Dr. Andreas Kerren, Associate Professor

Abstract
Social analysis is an important component of the field of Visual Analytics to support a better understanding of social spaces by visualizing relationships among people in chat rooms, forums, wiki-style spaces, and other social networks.

The development of tools and methods to synthesize information from massive, dynamic, ambiguous, and often conflicting data is a challenge that has been receiving growing attention on the last years. To address this challenge, researchers engage in visual analytics to organize information, generate overviews and explore the information space in order to extract potentially useful information.

This seminar paper presents and discusses new research approaches in social analysis and interaction. Several works are presented in four different articles selected from the last two editions (2006 and 2007) of the IEEE Symposium on Visual Analytics Science and Technology. Selected submissions span important visual analytics topics such as authors collaboration in the presence of controversy, conflicts and coordination costs in web-based collaborative authoring environments, and how to explore connections and eliminate duplicates on bibliographic collaboration networks. In this work, I will primarily focus on interaction techniques and visualization tools, how these tools works, how they enhance the human insight into abstract data and how these tools are applicable in social networks.

Keywords: *visual analytics, wiki, revert, graph, collaboration, user model, visualization, computer-supported cooperative work*

Table of Contents

1. Introduction

1.1 Introduction

Visual analytics is the formation of abstract visual metaphors in combination with a human interaction. Social analysis is an important component of the field of Visual Analytics and can be an important instrument to synthesize information from massive, dynamic, ambiguous and often-conflicting data. It helps us to identify patterns that are extremely difficult to express and discover purely analytically." Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis" [5].

As our understanding of social human interaction is very limited, it is a challenge to create well-constructed visual representations. We may have a concept of what is in that information space but no clue what is not there [5]. Visual representations of social network make it easy for users to perceive aspects of relationships among actors in a network and provide a better understanding of social processes and behaviours.

An increasing number of software tools are developed to help analysts in organizing their data to generate overviews and to explore the information space in order to extract potentially useful information. Most of these data analysis systems focus on data mining and interactive graph visualization and analysis [5].

Visualization tools have been developed for understanding conflicts and coordination costs in collaborative social spaces, such as Wikipedia, and to explore connections and eliminate duplicates on bibliographic collaboration networks. These tools try to provide answers on questions like: "Who revises whom?", "Who's connected to who?", "Who plays which role?" or "Who's collaborate to who?"

1.2 Goals of the Study

This report is a study of four papers selected from the last two editions (2006 and 2007) IEEE Symposium on Visual Analytics Science and Technology. The aim of this study is to present an overview of analysis and visualization techniques that reveal how authors collaborate in the presence of controversy and conflict patterns among groups of users in Wikipedia and authors' connections on bibliographic collaboration networks. In addition, the study offers insights on design considerations for asynchronous collaboration in visual analysis, environments, especially works parallelization, communication, and social organization.

1.3 Delimitations

In this paper, I will primarily show how these tools works, how they enhance the human insight into abstract data and how these tools are applicable in social network. I do not deeply go into technical questions and hade not the possibility to test all interaction tools on described-system.

2. “Who revises whom” in Wikipedia

Wikipedia is a large web-based collaborative authoring environment where anyone on the Internet can create, edit, and delete pages. Ward Cunningham, who launched the first wiki in 1995, introduced the term “wiki” [9]. Wikipedia is currently the largest wiki, maybe due to its size, popularity, and relevance for understanding new forms of collective knowledge creation; Wikipedia receives increasing interest in research. Moreover, the fast growth of Wikipedia presents a challenge for analysts to understand conflicts and other social dynamics. To address this challenge, Brandes and Lerner from the Department of Computer & Information Science, University of Konstanz, have elaborated a model of how conflicts occur in Wikipedia and how conflicts are resolved.

Based on idea that controversy is reflected in the reply behaviour of authors (revision behaviour in this case) [1], Brandes and Lerner present several improvements: their approach reveals authors´ involvement and roles, instead of dividing authors in opinion groups. Their visual analytics approach reveals rich insights into controversies, including who are the dominant authors, what roles they play, and how they interact.They also provide tools to understand how Wikipedia authors collaborate in the presence of controversy [2].

2.1 Method

The authors were interested in how Wikipedia authors collaborate when writing about controversial topics (such as abortion, gun rights vs. gun control etc) delicate historic events or important political persons. Such pages have often been revised up to tens of thousands times by several thousand authors who, possibly, not all share the same opinion on the particular topic. For example, the most-revised page in the English Wikipedia is George W. Bush having 33,086 revisions and 10,167 different authors (registered or anonymous) in December 2006 [2].

Wikipedia makes its complete database (containing all versions of every article since its initial creation) available in XML-format. The files containing the complete history of all pages can be extremely large; the complete dump for the English Wikipedia unpacks to more than 600 gigabytes (GB)[1]. For this study, the authors used so-called stub-file for the English Wikipedia from the 20061130 dump with a size of 23 GB [2]. A stub-file contains meta-data about every revision but not the text (see Figure 1).

An important stage to develop an efficient method for analyzing interaction among Wikipedia authors is defining so called “who-revises-whom”-network (in short “revision network”). A revert refers to a situation in which a user changes an article back to a previously written version, casting out changes that have been made [2]. Any work done on the article since the revert (including the revert itself) is lost.

The authors define a revision or edit to be a tuple of the form

r = (page, time, author, comment, revert)

where:

- *page* is a text-string denoting the page-title;
- *time* contains the exact timestamp of the revision (given by the second);
- *author* is a real user name if the contributor of the revision has been logged in or an IP-address if the revision has been done anonymously;

[1] http://meta.wikimedia.org./wiki/Data.dumps[2].

- *comment* is free text explaining what has been done or why this revision has been necessary; and *revert* is a Boolean flag labeling the revision.

```
<page><title>Gun politics</title>
...
<revision><timestamp>2006-03-18T22:31:41Z</timestamp>
<contributor><ip>24.12.208.181</ip></contributor>
<comment>/* Self-defense */</comment>
</revision>
<revision><timestamp>2006-03-18T23:18:38Z</timestamp>
<contributor><username>Yaf</username></contributor>
<comment>rv POV edit (discussion belongs on discussion page,
not in article)</comment>
</revision>
<revision><timestamp>2006-03-19T02:39:25Z</timestamp>
<contributor><ip>24.12.208.181</ip></contributor>
<comment>/* General discussion of arguments */ Fact with cite.
DO NOT DELETE WITHOUT VERY GOOD REASON!!!!!!!
Different placement on page acceptable.</comment>
</revision>
<revision><timestamp>2006-03-19T02:52:41Z</timestamp>
<contributor><username>Mmx1</username></contributor>
<comment>wikipedia is not a collection of facts. This page
is a summary of the arguments, not a place to make them</comment>
</revision>
<revision><timestamp>2006-03-19T05:24:30Z</timestamp>
<contributor><ip>24.12.208.181</ip></contributor>
<comment>HUH?? Facts don't belong in this article.
Can that be true?</comment>
```

Figure 1: Six consecutive revisions of the page Gun politics in XML format (picture taken from [4]).

Given a sequence $R = (r_1, \ldots, r_N)$ of revisions on the same page, which is ordered by increasing timestamps, the associated revision network is a directed, weighted graph $G = (V, E, \omega)$ defined as follows (also compare Figure 2):

- $\mathbf{V}$ is the set of authors that performed a revision in R.
- $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ is the set of revision edges. For two different authors $\mathbf{u}, \mathbf{v} \in \mathbf{V}$ the edge $(u,v) \in E$ is introduced if there are two consecutive revisions $\mathbf{r}_i, \mathbf{r}_{i+1} \in R$ such that **u** is the author of $\mathbf{r}_{i+1}$ and v the author of $\mathbf{r}_i$. An edge **(u,v)** can be read as "u revises changes made by v".
- The function $\omega: R \to \mathbb{R}$ assigns weights to edges. For an edge (u, v) the weight $\omega(u, v)$ indicates how "urgent" u considers it to revise the changes made by v.

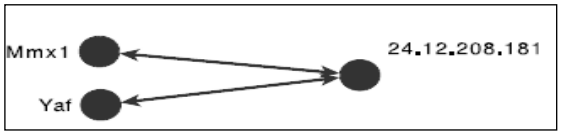

Figure 2: Revision network arising from the six revisions shown in Figure 1 (picture taken from [2]).

2.2 Visual Analysis of the Revision Network

Another important step is to define and explain the features of the authors, and graphically represent them. These characteristics include the authors' position, their involvement in controversy, an indicator showing if they are mostly revisors or mostly being revised, and an

indicator telling whether their edit behaviour is rather constant over time or highly concentrated on shorts periods. Technically, the determination of the authors' positions it is a complex task [2].

The position of a particular author should express which other authors he/she confronts. Confrontation is reflecting in the revision edges: if two authors take different positions, they disagree with the edits of the other and therefore will frequently revise each other. Thus, if two authors **u** and **v** are connected by a revision edge of large weight, then **u** and **v** are drawing on opposite sides with the whole network, thus all confronting pairs are simultaneously as far from each other as possible.

To efficiently solve this problem they associate a revision network with author set V of cardinality n = |V| with its *symmetric adjacency matrix* A = (a_{uv}) with rows and columns indexed by V and entries $a_{uv} = \omega(u,v) + \omega(v,u)$ corresponding to the sum of the weights of the two directed edges between the two endpoints.

The following algorithm is used for determining the authors' positions and involvement. It takes as input the symmetric *adjacency matrix A* of the revision network [2].

- Compute the smallest and second smallest eigenvalue lmin and lm0 of A and the associated (normalized and orthogonal) eigenvectors x and y.
- Set s=lm0 in/lmin as the network's skewness and define for an author v its position p(v) = (p1(v), p2(v)) = (xv, s · yv) 2 R2 and its involvement i(v) =pp1(v)2+ p2(v)2.

2.3 Visual Representation

To visualize a revision network the authors proposed a variant of spectral graph clustering heuristics. Visualizing the complete revision network over the whole lifetime of the page gives an overview revealing the most important authors, the roles they play, and the other authors they confront. To determine the relevant sub-structures of the revision network, the information have been filtered with restriction to time intervals and restriction to relevant sub-networks.

The edit volume diagram shown at the bottom of the images reveals time points when the page receives much interest. It is straightforward to restrict the revision network by including only revisions within a certain time interval (see Figure 4).

The goal of the network clustering is to put authors who strongly revise each other into the same cluster, and authors that have only little interaction into different clusters. The sub-networks induced by the various clusters are then analysed separately to identify recurrent patterns of confrontation.

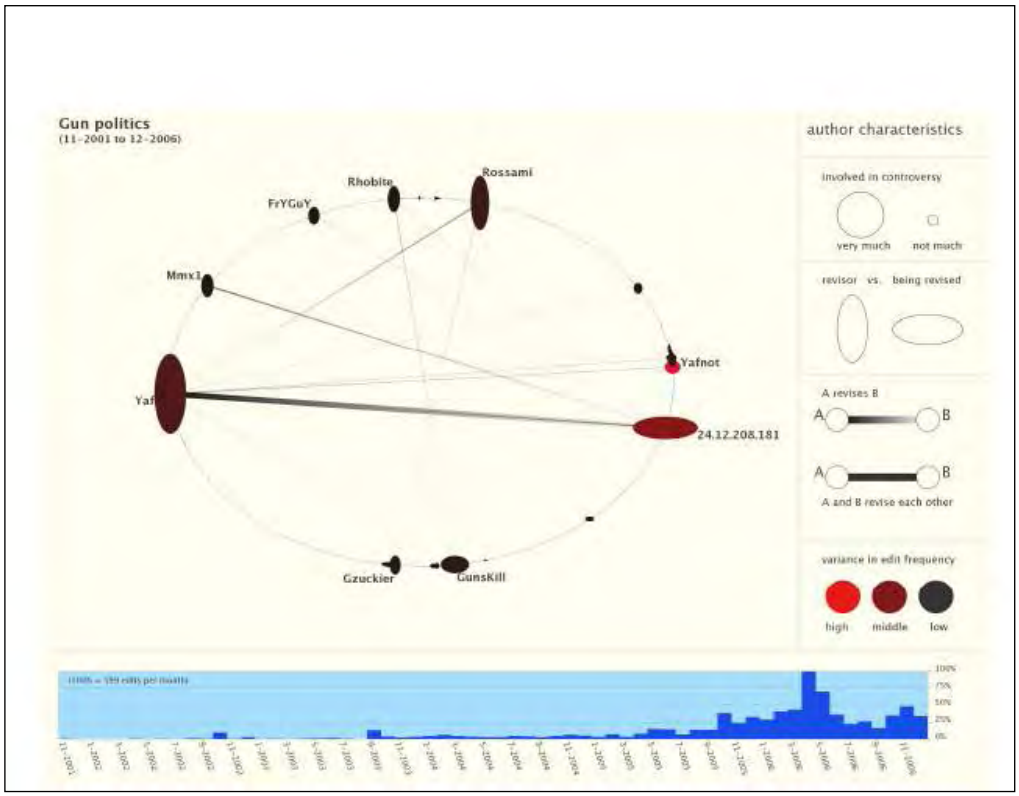

Figure 4: This screenshot illustrates the visualization of a revision network determined from Gun politics and related pages. Nodes represent the different authors. If two authors are on opposite sides, they strongly revise each other. Other characteristics such as how much the authors are involved in controversy, revisor vs. being revised, and variance in edit frequency are represented in the legend on the right-hand side. The diagram at the bottom shows the total number of edits per month (picture taken from [2]).

This example illustrates the confrontation between authors with different points of view about carrying guns. The dominant confrontation is clearly between "Yaf" and the anonymous user "24.12.208.181". "Yaf" advocates the freedom to carry guns and the other takes the opposite point of view. Besides differences in opinion, another distinction between these two users is that "Yaf" is more a revisor and "181" more revised. Author "Yafnot" is an example of a user that did not contribute much (only seven edits) but is quite a lot involved in controversy
Looking at the sequence of edits , taking into account the positions of "Yaf", "Yafnot", and "181", and considering the purposeful name of "Yafnot", could come to the hypothesis that "Yafnot" and "181" are the same person. The supposition was true and the user Yafnot was blocked on 2'nd April 2006 (still less than two hours after his/her first edit) by Rhobite for impersonation [2].

2.4 Results

Concrete contributions of this research consist on techniques for visual analysis of the revision network, which reveal the authors that are the most involved in controversy. This approach shows "who confronts whom" and "who plays which role" and offer solutions to identify some recurrent patterns of confrontation.
The benefits of the presented visualization lie in the fact that this approach can be applied to Wikipedia articles in any language without the need for adapting *NLP algorithms*[2]. This is a significant advantage since for most languages text-processing algorithms are not so highly

[2] Natural Language Processing Algorithms

developed as for English. It is important to note that the analysis cannot and does not attempt to determine which opinion is more acceptable.
One issue for future work is to determine more conclusively the difference between opinion-triggered and vandalism-triggered confrontation otherwise, interpretation of the revisor vs. revised pattern can be quite different. Another issue is to improve the construction of the revision network by taking into account whose text was been changed during a revision.

3. "Us versus them" in Wikipedia

Since its foundation, Wikipedia has been growing with an exponential rate [5, 6, 18, 32, 47]. Over 2 million articles have been collaboratively edited by more than 4 million users in the English Wikipedia only [10]. In fact, huge investments in time and money are often lost, because we still lack the possibilities to analyse and visualize collaborative spaces. However, the high level of participation in growing organization comes with corresponding costs.

Starting from the same necessity of analyzing disagreement inside this social network, Suh, Chi, Pendleton and Kittur from Palo Alto Research Center consider that some of these coordination costs result from user disagreements about article content, procedures, and administrative issues. They offer a modality for identifying patterns of conflicts in Wikipedia articles.

3.1 Method

To understand disagreements among users, the authors build a model of how users engage in disputes. The model, called "user conflict model", is based on user's editing history and the relationships between user's edits, especially revisions that void previous edits, known as "reverts" [10].
Reverts are often use to fight against vandalism and to bring articles back to their original state. However, users also use reverts to block other users' contributions. "Edit wars" [10] are a typical example where disagreeing users repeatedly revert each other's edits. Assuming that reverts are proxies for dispute and disagreement, the group identify reverts in Wikipedia by two different methods:

- Data-driven method uses a unique identifier of every revision made to every article using the *MD5 hashing scheme*. [3] The hashing functions create a small fingerprint of each revision, which is suitable for rapidly comparing all revisions of an article. Using MD5 values for all revisions of an article, makes an identification possible when a later revision exactly matched the hash of a previous article, indicating a revert.
  The advantage of this method is that it does not depend on users to label reverts. The disadvantage of this method is that it does not pick up partial reverts, in which only some of the text in an article is reverted.
- User-labelled method to capture partial reverts including revisions whose revision comments included the text "revert" or a commonly used abbreviation of revert "rv".

The combination of both methods provides converging evidence on the true change in reverts over time. Figure 5 illustrates that the statistics for reverts calculated by the two methods have slightly different characteristics.

[3] MD5 hashing scheme is commonly used to check that data objects are identical.

Reverts (MD5 hash method)	3,711,638
Self-reverts	582,373
Pages with at least one revert	721,866
Pages with 50 reverts or more	9,973
Reverts (Comment method)	2,422,482
Vandalism(Comment with vandal, rvv, etc)	577,642
Reverts (Union of both methods)	3,917,008

Table 1: Revert and Vandalism Statistics (picture taken from [10]).

According to this user conflict model they extracts reverts from Wikipedia editing history and composes a node-link graph where a user is denoted as a node and a revert relationship as a link. However, using reverts to identify conflicts is hard because:

- multiple users are often involved in chains of reverts
- edit history is typically long and tedious to browse
- various types of reverts
  - the "revert duel"
  - the "self-reverts"
  - reverts by multiple users

To solve this problem a number of design choices was taken in consideration:

- *Disregard Self Revert.* Self-reverts are disregarded.
- *Degree of Conflict.* The measuring of the amount of dispute between two users is given by the number of reverts between them.
- *Conflict Group.* When two users make reverts on edits made by another user, but not against each other, the two users are presumed to have similar opinions.
- *Identity Based Revert.* The MD5 method is used to identify reverts. When two revisions have the same textual content, they define the later edit as revert.
- *Immediate Revert Only*. When an article page is reverted to an older version other than its immediate last version, the intention of the revert is ambiguous because it is not clear whether the revert is exclusively toward the last edit.

The authors implemented this approach using a *force-directed graph layout algorithm* that assigns forces in a way that the edges (representing revert relationships) act as springs, while the individual users are represented as particles with gravitational fields. Users (represented as nodes) attract each other unless they have a revert relationship. A revert relationship is represented as an edge, thus pushing such users apart (as shown in Figure 6).

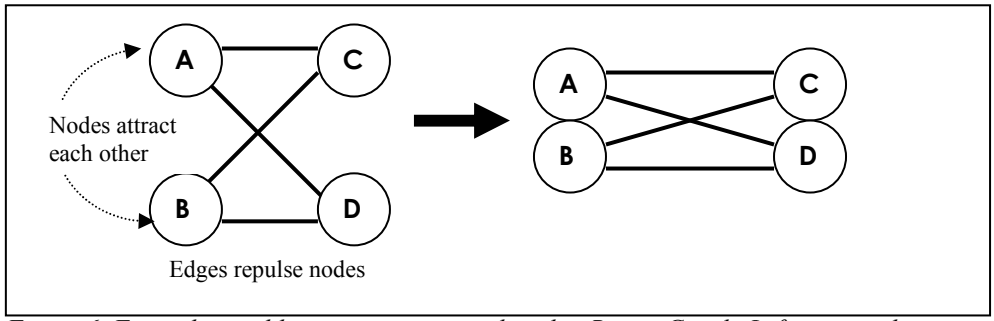

Figure 6. Force directed layout structure employed in Revert Graph. Left part: nodes are evenly distributed as an initial layout. Right part: when forces are deployed, nodes are rearranged in two user groups (picture taken from [10]).

3.2 Visualization and Patterns

Based on the user model, they developed a visualization tool called "Revert Graph". The graph enables visual analysis of opinion groups and rapid interactive exploration of those relationships via detail drilldowns. Suppose, a user wants to investigate conflicts and disagreements inside a Wikipedia article. The tool allows the user to specify an article she/he wants to explore by typing the name of the article. Then, the revert history of the article is retrieved from a database and a node-link graph is formed and displayed on the screen. A force-directed layout module then clusters user nodes based on revert relationships. Figure 7 shows an example.

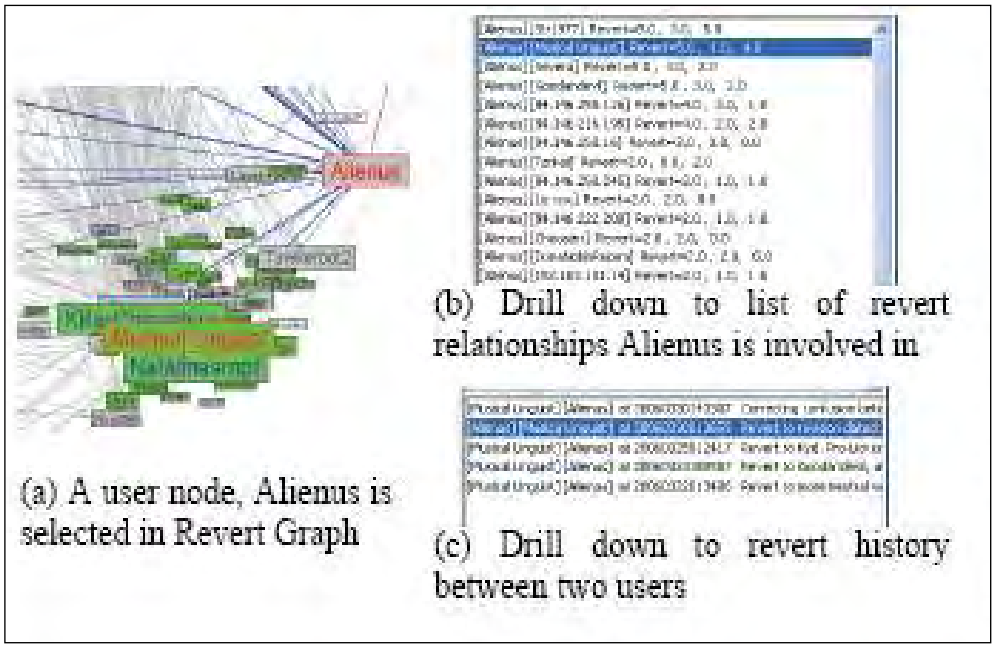

Figure 7. Enlarged view of the Terri Shiavo page in Revert Graph (picture taken from [10]).

The tool can answer questions such as the severity and form of the disagreement as well as the shape and size of opinion groups. Revert Graph also provides ways to change zooming level, node size, and other visual options. It was designed to help identify user groups representing opinion groups, the specific motivation of revisions, and the conflict detail.
To examine a potential user conflict pattern in an article, the analysis involved detail investigation of the article revision history. To get a more clear insight on users' position on the issues of an article, the authors browsed through information such as revert comments, article talk pages, user pages, and users' edits on other pages is used.

To demonstrate the effectiveness of the tool sub et al. selected 901 high conflict articles with more than 250 reverts for analysis. These articles contain a large amount of discussion with extensive editing history. They find and pinpoint patterns, such as:

1. **Formation of opinion groups patterns**
   To obtain users' points of view on the topic, they browsed their user pages, user talk pages, revision histories, revision comments, as well as specific reverts.
2. **Mediation patterns**
   Another common pattern revealed by Revert Graph is a group of users attempting to mediate among user groups with divergent points of view.
3. **Fighting vandalism patterns**

Revert Graph uncovers clear patterns of vandalism and anti-vandalism efforts. It was found that an anti-vandalism robot Tawkerbot2 [8] is often actively engaged in this pattern.

4. **Controversial editors patterns**
   For example, Figure 8 shows how certain controversial editors are easily identifiable in the Revert Graph by the size of a user node and the thickness of the edges representing the degree of revert relationship between users. By these means, the visual saliency of editors engaged in many conflicting reverts is increased so these users are more quickly identified. They are usually self-appointed experts, or have strong points of view.

The Wikipedia page on Dokdo is one example where they find interesting formation of opinion groups' patterns. Dokdo is a disputed isle in the Sea of Japan (East Sea) currently controlled by South Korea, but also claimed by Japan as "Takeshima". Figure 8 shows opinion groups discovered on the Dokdo article [11].

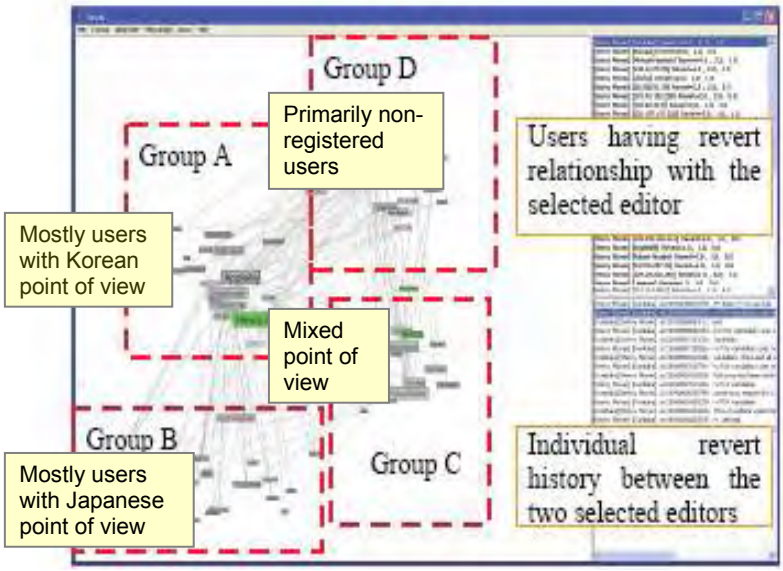

Figure 8: Revert Graph for the Wikipedia page on Dokdo (picture taken from [10]).

As shown in this conflict pattern, node clusters identified in Revert Graph correspond to opinion groups that are not only cohesive but also represent major points of view in these topic areas.

Number of users in user group	A	B	C	Total
Users with Korean point of view	10	6	0	16
Users with Japanese point of view	1	8	7	16
Neutral or Unidentified	7	3	6	17

Table 2. User Groups on the Dokdo article.

The analysis is summarized in Table 2 shows that the identified user groups indeed represent distinct opinion groups. Inside Group A the number of users with Korean point of view are higher than users with Korean point of view, inside Group B the users with Japanese point of view are higher than users with Korean point of view. In the graphical representation, the user node for Group A is bigger those from Groups B and C.

3.3 Results

Growing organization often requires overhead costs, such as coordination and maintenance work Wikipedia is not an exception. Some of these coordination costs result from user disagreements about article content, procedures, and administrative issues. Research needs to understand how visual analytics can help analysts to understand conflicts in collaborative spaces and their consequents.
The Revert Graph was usefully uses to identify important social patterns in Wikipedia. Moreover, the tool may be applicable on other wiki-based systems in which reverts are tracked as part of system usage [10].

Revert Graph offers a solution on development of conflict resolution tools. However, because it restraint requires sufficient revert relationships in the data set, not every aspect of social dynamics is fully addressed. For instance, it cannot detect conflicts between users who were not involved in reverts [10].

4. ”Who's connected to who” on Bibliographic Collaboration Networks

The wiki-based systems are not the only ones in need for tools that can help visualizing and analysing social networks.Visual analytics can also provide useful tools for users to make sense of complex collaborative environments such as the bibliographic domain.
An important first step, before analysis can begin, is ensuring that the data is accurate when dealing with a diverse collection of information selected from databases. In the bibliographic domain common errors that lead to duplicates in databases are: 1) parsing errors, such as switching a first name and last name, 2) abbreviations, such as using first initial instead of full first name, and, of course, 3) misspellings.[3]

Defining the problem, the data may inadvertently contain several distinct references to the same underlying entity or actor. The previous work shows that calculating any of the standard social network measures, such as degree-centrality[4], betweenness[5] , closeness[6] and so on, would give inaccurate results that lead to:

- Visual display is misleading: incorrect number of nodes & the edges and paths are inaccurate.
- Calculating of the standard social network measures would give inaccurate results [3].

The work by Bilgic at el. describes a novel approach for an application of visual analytics techniques in social collaboration networks, in particular using entity-resolution. The authors illustrate the benefits by trying to identify potential duplicates of authors in the bibliographic database.

Using this method in social networks is more interesting because the social context, or “who's connected to who”, can provide useful information to the resolution process. Innovatively, the

[4] The count of the number of ties to other actors in the network [10].
[5] Degree an individual lies between other individuals in the network [10].
[6] The degree an individual is near all other individuals in a network (directly or indirectly) [10].

contributions of the work include providing an intuitive and directly accessible representation of data.

4.1 Method

Presently, the existing entity resolution methods use automated or hand-cleaning methods. Automated techniques are not perfect, and they face a precision-recall trade-off. If they are tuned to have high precision, they rarely merge duplicates, leaving many duplicates in the database. If they are tuned to have a high recall, they mistakenly merge nodes that are in fact distinct [3]. The hand cleaning methods can be slow and inefficient in finding duplicates. These approaches tend to be high precision, because there is a human-in-the-loop making the final resolution decision.

Bilgic at el. provide an interactive analyst-centric approach that integrates the *data mining techniques* with visualization appropriate to the task. They built an interactive tool, D-Dupe, to provide access to sophisticated entity resolution algorithms and enables users to apply sequences of actions to uncover duplicates. D-Dupe resolves ambiguities either by merging nodes or by marking them distinct. Figure 9 gives an overview of the deduplication process on a small portion of bibliographic dataset [3].

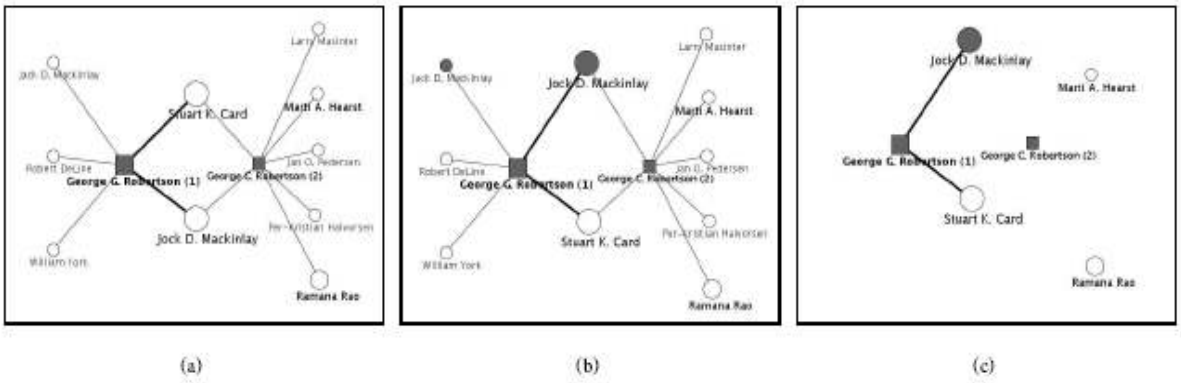

Figure 9 : (a) The initial collaboration network for potential duplicates George G. Robertson and George C. Robertson. (b) The use of the Co-Authorship Similarity Slider highlights another potential duplicate among the neighbors: Jack D. Mackinlay and Jock D. Mackinlay (c) Filtering the collaboration network using the node and edge weights quickly isolate George C. Robertson from the rest of the network signaling that it might be a misspelling (picture taken from [3]).

4.2 Visualization

D-Dupe, written in Java, integrates data mining algorithms with an interactive visualization interface.Two of D-Dupe's novelties are “a stable visual layout optimized for entity resolution and a user control for combining entity resolution algorithms” [3]. D-Dupe’s layout consists of three coordinated windows: (see Figure 10)

1. On the left, a window prove the collaboration context network panel
2. The entity resolution control panel on the right
3. The potential duplicates details panel at the bottom

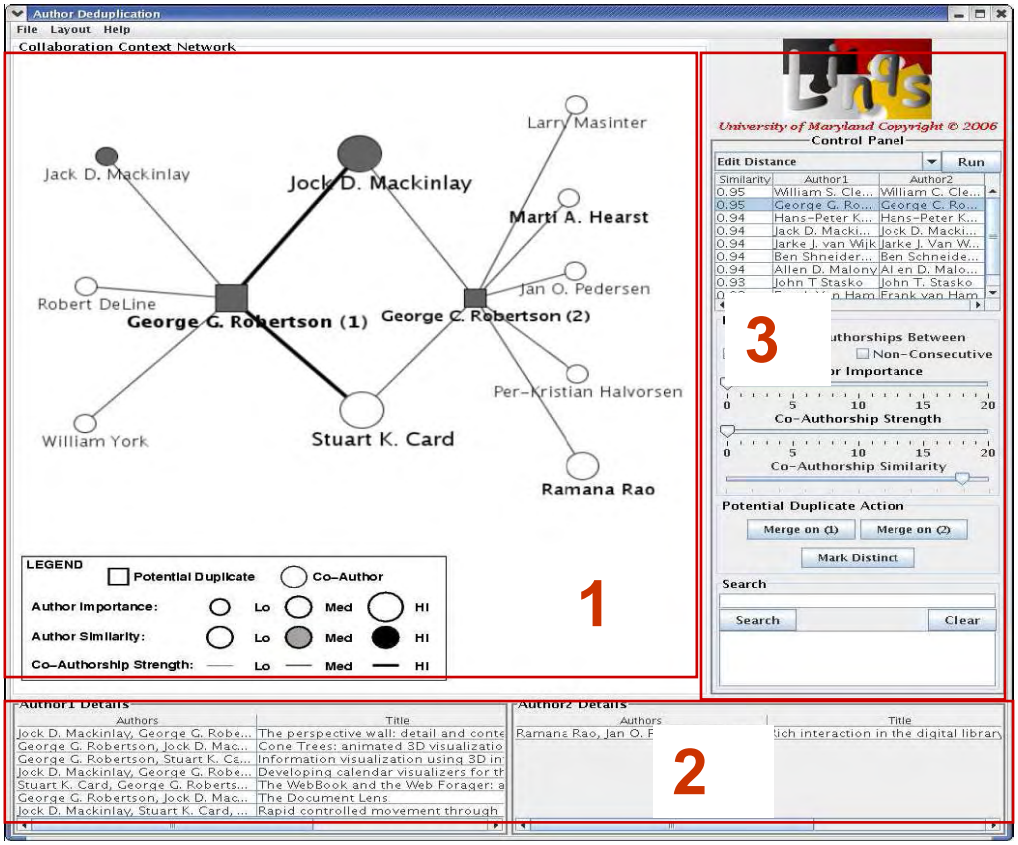

Figure 10: D-Dupe overview (picture taken from [3]).

Additional information about the nodes is conveyed through shape, shading, and size. The current potential duplicate nodes are squares, and the other nodes are circles. The current potential duplicate pair and other potential duplicates in the neighborhood are shaded according to their similarity based on the current entity resolution metric. Darker nodes indicate a greater degree of similarity. The similarity shading for the nodes in the neighborhood can be controlled using a slider.
The layout shows only the subnet work relevant for the entity resolution task and allows visualization to scale to large networks. An important aspect is that the potential duplicates and other related entities always appear at the same location

The user control allows flexibly applying and interleaving different measures. Numerous similarity measures can be used to determine potential duplicates. The user can select an entity similarity measure to use, view a list of candidate duplicate pairs, choose alters for the nodes, edges and collaborators, perform resolutions for a particular pair, and search for a particular author. D-Dupe allows dynamic filtering of the collaboration context network [3].
The interaction paradigm is as follows:

1. User start on loading a dataset.
2. Choose from a number of possible entity resolution algorithms.
3. The entity resolution algorithms ranks pairs of nodes according to how likely they are to be duplicates.
4. User selects a potential duplicate pair for analysis.
5. The tool views the collaboration context network for the pair and applies filtering and highlighting.
6. Finally, users have to decide that the two nodes are duplicates or distinct node.

User actions are recorded and at any point in the process, the 'resolute' network can be saved. Recording resolution allows users to examine history of the resolution decision. The resolution process is iterative [3].

4.3 Results

D-Dupe is a visual analytics tool created to resolve duplicates in different database and on three bibliographic datasets. D-Dupe's effectiveness was demonstrated in several projects, where it has been used to detect, highlight duplicates and to eliminate duplicates. For example, using D-Dupe for cleaning the "CiteSeer" dataset, 10 duplicates in 20 minutes was detected and resolved [3].

D-Dupe's layout and interaction principles are general and can be used in other social networks in which the relational context provides useful information for entity resolution decisions. Given that D-Dupe uses several standard string similarity functions, including Levenstein, Jaccard, JaccardChar, Jaro, JaroWinkler and MongeElkan , the users can select from a variety of entity similarity metrics to identify and rank potential duplicates.
D-Dupe illustrates the utility of building interactive tools that combine data mining and information visualization to support specific analytic tasks.

5. Design Considerations for Asynchronous Collaboration

The basic idea of visual analytics is visually represent the information, allowing the human directly interact with the information, to draw conclusions, and to make better decisions. Most of researches assume a single-user focus on perceptual and cognitive processes, in practice, the sensemaking process is often a social one and should use approaches that support social interaction.In addition, exploring large data sets from a single-user perception is difficult and inadequate. This suggests that suitable interactive visualization tools should also support social interaction.
Heer and Agrawala from University of California provide a guide for design and evaluation of collaborative visualization systems to fully support sense-making model. Previously, they began exploring this area by building and evaluating *sense.us*, a system for *asynchronous collaborative visualization* (see Figure 11). By partitioning work across both time and space, asynchronous collaboration offers better scalability for group-oriented analysis [3].

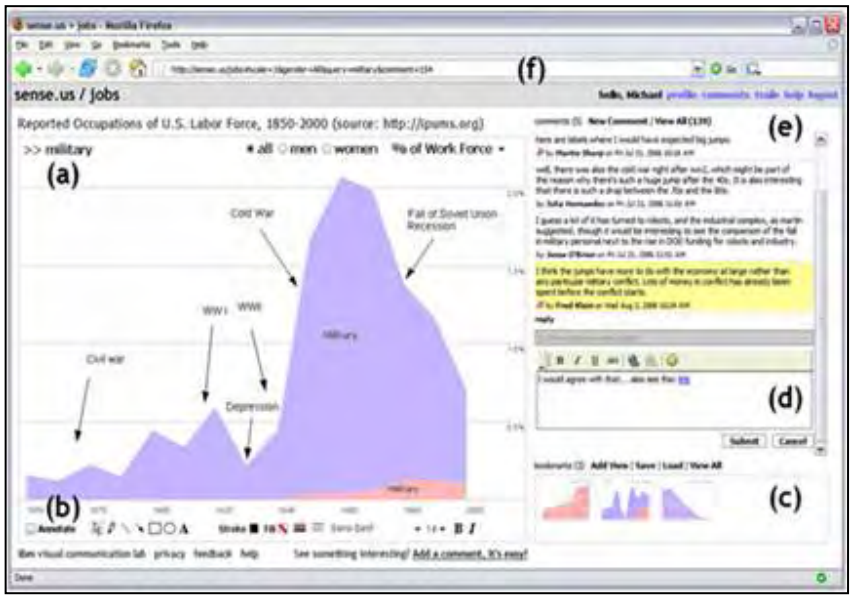

Figure 11:Sense.us (picture taken from [6]).

Based on the observations, they discover numerous examples of group sense making in action: question, hypothesis, identification of problematic or incorrect data values and social navigation to interesting or controversial data. Wanting a better support for these observed behaviours, they suggest design decisions in both theoretical and practical knowledge of group interaction [4].

Regardless of the fact that the most appropriate collaboration mechanisms for supporting social interaction are not immediately clear, creating effective mediated collaboration environments raises a number of design questions:

- How should collaboration be structured?
- What shared artifacts can be use to coordinate contributions?
- What are the most effective communication mechanisms?

5.1 A Set of Design Considerations

For a better understanding of analytic contributions, the authors consulted the sense-making model [9], which grounds the use of information visualization in a theory of how people search for, organize, and create new knowledge from source information [4]. They study asynchronous collaborative visualization systems that support collaborative analysis around both statistical and geographic data. These systems support varied levels of sharing, discussion, and annotation of visualized data; each supports simple text comments and view sharing through book marking. The systems are: Spotfire Decision Site Posters, WikiMapia, Swivel, Sense. us, and Many Eyes.
WikiMapia is an online map and satellite imaging resource that combines Google Maps with a wiki system, allowing users to add information (in the form of a note) to any location on earth.
Many Eyes is a live service that allows users to upload their own data sets and create visualizations of them based on a series of graph and chart templates provided by the site (see Figure 12).

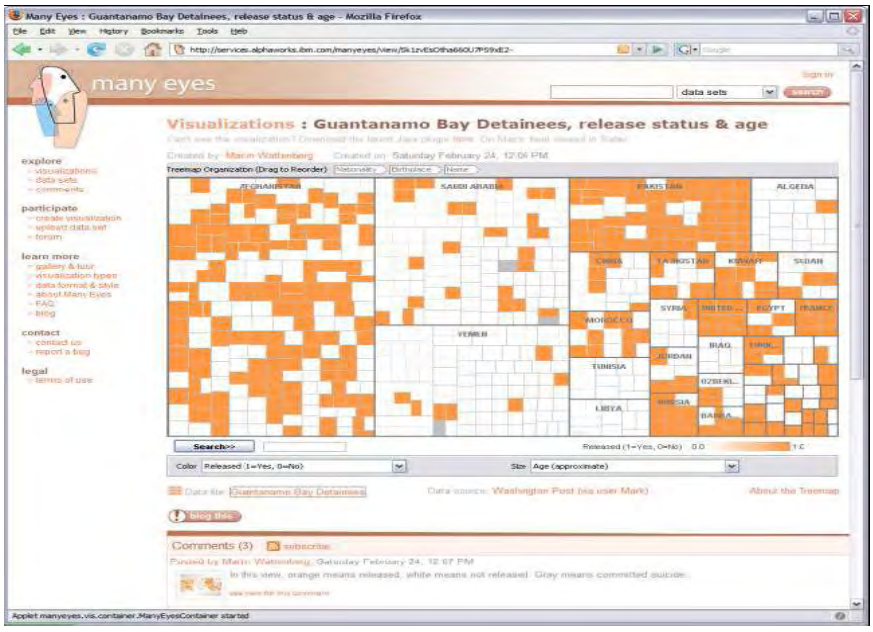

Figure 12: Many Eyes (picture taken from [6]).

They identify a set of design considerations grouped into seven topical areas. In each of these areas, they discuss the fundamental *activities* that enable effective collaboration, and suggest specific *mechanisms* by which they could achieved [4].

1. Division and allocation of work
2. Common ground and awareness
3. Reference and deixis
4. Incentives and engagement
5. Identity, trust, and reputation
6. Group dynamics
7. Consensus and decision making

We explain now each mechanism in detail.

5.1.1. Division and allocation of work

An important aspect of collaborative visualization is how to facilitate the modularization[7] of work. The first step is determining the modules of work and their granularity[8]. Primary concerns are how to split work among multiple participants and significantly combined the results and the allocation of individuals to tasks in a manner that best matches their skills and disposition. To identify the module of contribution they use a general pattern for describing visualization applications, the *information visualization reference model* (see Figure 13).

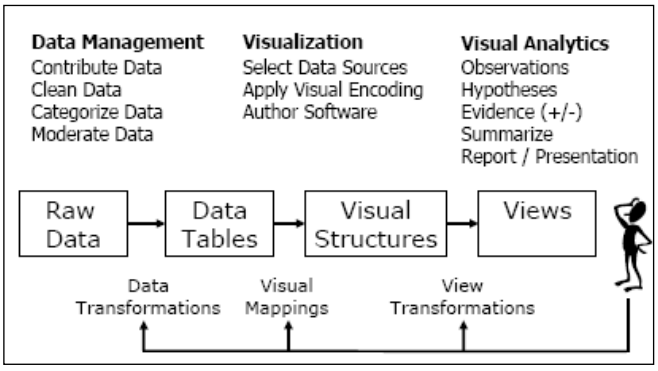

Figure 13: The information visualization reference model (picture taken from [6]).

Moreover, to understand how people search for, organize, and create new knowledge from source information they consult the so-called sensemaking model (see Figure 14).

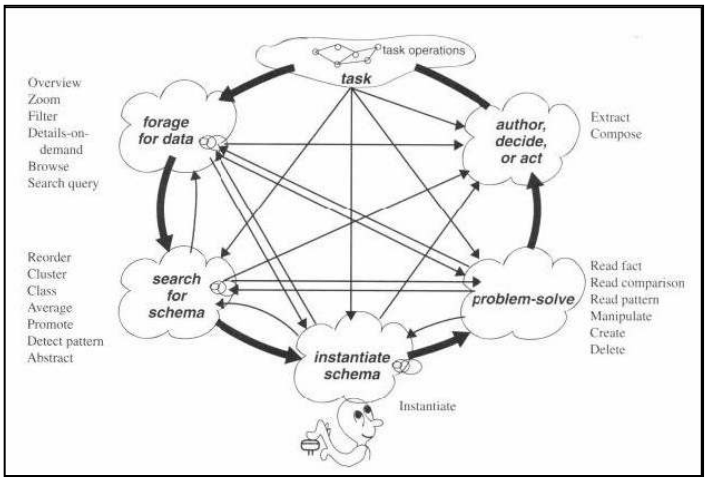

Figure 14: The sensemaking model(picture taken from [4][5]).

[7] *Modularity* refers to how work is segmented into atomic units, parallelizing work into independent tasks [4].
[8] The *granularity* of a module is a measure of the cost or effort involved in performing the task [4].

The *information visualization reference model* separate the visualization process into data acquisition and representation, visual encoding of data, and display and interaction. Each segment of this model provides an entry point for collaborative activity such as uploading data sets, cleaning or reformatting data, moderating contributed data, affixing metadata.
Many Eyes and Swivel are example of systems that enable contribution of data sets and visual mappings. These systems support varied levels of sharing, discussion, and annotation of visualized data. The *sensemaking process* has a much higher degree of coupling than the information visualization reference model, carrying implications for the granularity and integration of contributions.

5.1.2. Common ground and awareness

Another approach is the *embedded discussion*, placing conversational markers directly within the visualization, such as comments over annotated geographic regions in WikiMapia. The approach provides unidirectional links that point from the visualization to text. This form of *independent discussion* is unidirectional, linking from text to the visualization. Independent, unthreaded comments are used in Many Eyes.

5.1.3. Reference and deixis

An important design consideration is reference to artefacts, people, places, or other items. For improving collaboration, it is important to establish and grounding the reference between participants to eliminate ambiguity of reference and understand how various forms of reference may be applied. For example, in collaboration around visual media common reference may be *general* (e.g., east by northeast), *definite* (e.g., named entities), *detailed* (e.g., described by attributes, such as the red rose), or *deictic* (e.g., pointing to an object and saying that one, also referred to as *indexical* reference).
Forms of spatial indexical reference are grouped into categories of *pointing* and *placing*. *Pointing* behaviours use some form of vectorial reference to direct attention to an object, group, or region of interest, such as pointing a finger or directing one's gaze. The deictic pointing gestures can play an important role and state that successfully supporting deixis can improve visualization techniques. Providing interaction techniques for pointing designers might not only aid human communication, but also allow for machine-readable forms of pointing or annotation, supporting a navigable index of references [4].

Placing behaviours involve moving an object to a region of space that has a shared, conventional meaning. Another design consideration is how various forms of reference may be applied in tandem. For example, one might deictically refer to a particular object, but formulate a broader selection by abstracting from the properties of that object (e.g., select all items that are blue like this on). The implicit interplay between gesture and text, often segmented in time and interpreted subconsciously in synchronous interactions, may need to be more concretely reified in asynchronous contexts.

5.1.4. Incentives and engagement

In collaborative work, where professionals collaborators are involved in a particular context, there may be existing incentive, for conducting work.

Incorporating incentives into the design process may increase the quantity and/or quality of contributions. Design considerations for improve contribution rates consist on monetary incentives[9], hedonic incentives[10], and social-psychological incentives[11].

[9] *Monetary* incentives refer to material compensation, such as a salary or cash reward [6].
[10] *Hedonic* incentives refer to well-being or engagement experienced intrinsically in the work [6].

The social-psychological incentives can improve contribution rates by prominently display new discoveries or successful responses to open questions. Mechanisms for positive feedback, such as voting for interesting comments, might also foster more contributions.
One challenge for design is to consider what pieces of information are most informative for reputation formation. Some systems provide *explicit* reputation mechanisms, such as seller ratings in online markets (e.g., eBay).Other systems instead provide *implicit* means of reputation formation, allowing collaborators to make inter-personal judgements grounded in past activity.

5.1.5 Identity, trust, and reputation

In a computer-mediated environment, aspects of identity, reputation, and trust influence the way people interact with each other. Context of deployment is an important aspect considering implication of identity. If collaborators are already familiar to each other, it may be enough to simply identify collaborators' individual contributions with recognizable names. But if collaborators begin as strangers, mechanisms for self-presentation and reputation formation need to be included in the system design (e.g., identity markers, demographic profiles and group memberships).

Observations of social use of visualization have noted an affinity of visualization users for data that they find personally relevant. Selecting data sets or designing their presentation, thus that the data is seen as personally relevant, usage rates will rise due to increased hedonic incentive. An example are geographic visualizations that facilitate navigation to personally relevant locations through typing in specific zip codes or city names.

5.1.6. Group dynamics

The makeup of collaborative groups is another aspect important to social sensemaking. Issues, such as the choice of group size, the diversity of group members and group management mechanism, can be used to improve design.

Formal *group management* mechanisms present useful means for addressing issues of scalability and privacy. Groups provide a mean of filtering contributions, improving tractability and reducing information overload for participants who may not be interested in the contributions of strangers.Finally, groups provide a means of limiting contribution visibility, using a mechanism for individual privacy within large-scale online scenarios.

5.1.7. Consensus and decision-making

Considering group consensus is an important part in many steps of the sensemaking cycle, the authors consider implications of the discussion model and vote a common ground. Consensus may arise through discussion or may involve the aggregation of individual decisions. Discussions are an important dimension of group consensus and have to support commentary.
Quantitative measures can be used for consensus and to lower integration costs.
Common forms of information exchange in group sense-making are reports and presentations. The challenge of collaborative visualization is to provide mechanisms to aid the creation and distribution of presentations [6]. For example on WikiMapia, users can vote on the accuracy of labelled geographic regions.

---

[11] *Social-psychological* incentives involve perceived benefits such as increased status or social capital [6].

5.2 Discussion

Heer and Agrawala suggest that appropriate collaboration mechanisms are not well understood, and they develop considerations for guiding the design and evaluation of collaborative visualization systems. The overarching goal is to effectively parallelizing work, facilitate mutual understanding, and reduce the costs of collaborative tasks [6]. Design elements are used as key issues to guide work in collaborative visualization. By partitioning work across both time and space, asynchronous collaboration offers greater scalability for group-oriented analysis. There is evidence that, due in part to a greater division of labor, asynchronous decision making can result in higher-quality outcomes—broader discussions, more complete reports, and longer solutions—than face-to-face collaboration.

6. Conclusions and Discussions

Human existence depends on collaborative problem solving. It is therefore not surprising that Visual Analytics must work effectively in collaborative environments. The need for Visual Analytics was determined by a growing amount of data to analyze; increasing complexity and uncertainty in the data; a lack of methods, technology, or tools [5]. One of the challenges of visual analytics is data representation because data, in raw form, are rarely appropriate for direct analysis.

Visualization tools and design suggestions presented in this seminar work offer potential solutions for understanding collaborative social spaces. These are only a few of the many examples of challenges facing Visual Analytics. As conflict and coordination costs increase in such collaborative environments, Visual Analytic tools may be increasingly useful for users to make sense of the status of the collaborative environment.

This seminar work presents and discusses hot research topics; they help a human analyst to understand some data and underlying phenomena. Visual Analytics tools discussed in this paper are applicable to diverse social networks. As collaborative authoring environments become more common, analytical tools, such as Revet Graph or D-Dupe ,will be important.

7. References

[1] Agrawal R., Rajagopalan S., Srikant R. and Xu Y. (2003)*Mining newsgroups using networks arising from social behavior*. In Proc. 12th Intl. Conf. WWW, pages 529–535.

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

[3] Bilgic M., Licamele L., Getoor L., Shneiderman B. (2006), *D-Dupe: An Interactive Tool for Entity Resolution in Social Networks* ,Proceeding of IEEE Symposium on Visual Analytics Science and Technology 2006.

[4] Card, S.K., Mackinlay, J.D., Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision To Think*. Morgan-Kaufmann.

[5] Heer, J., Agrawala, M. (2006). *Software Design Patterns for Information Visualization*. IEEE Transactions on Visualization and Computer Graphics. 12(5). Sep/Oct 2006.

[6] Heer J. and Agrawala M.,(2007) *Design Considerations for Collaborative Visual Analytics* ,Proceeding of IEEE Symposium on Visual Analytics Science and Technology 2007

[7] Keim D. A., Mansmann F., Schneidewind J., and Ziegler H. (2006) *Challenges in visual data analysis*. In IEEE Information Visualization, London, UK, 2006.

[8] Kittur A., Suh B., Chi, E. H., Pendleton, B. A. (2007) *He says, she says: Conflict and coordination in Wikipedia.* CHI 2007; 2007 April 28 -May 3; San Jose; CA.

[9] Leuf B. and Cunningham W. *The Wiki Way*. Addison-Wesley, 2001.

[10] Suh B.,Chi E., Pendleton B., and Kittur A.,(2007) *Us vs. Them:Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations*, Proceeding of IEEE Symposium on Visual Analytics Science and Technology 2007.

[11] Wikipedia.org, Dokdo. http://en.wikipedia.org/wiki/Dokdo 2008.

[12] Wikipedia.org, Network analysis. http://en.wikipedia.org/wiki/Network analysis 2008.