A Visualization Platform For Exploring Cooperation

Remy Cazabet
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, japan
remy.cazabet@gmail.com

Hideaki Takeda
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, japan
takeda@nii.ac.jp

ABSTRACT

In this paper, we present a platform designed to explore visually massive cooperation between individuals. With the increasing importance of the Internet, new types of cooperation are becoming common, in which hundreds, thousands or millions of individuals act together in interaction, and produces content in a decentralized manner. As these processes are happening in real-time and without organization, individuals involved in them often do not have a clear vision of what is happening, or even which role they play in it. The visualization we propose would allow users to take back the power of understanding the processes to which they participate in. We combine time series visualization, together with custom network visualization, in a way generic enough to adapt to many situations, while offering numerous possibilities.

1. INTRODUCTION

Since the advent of the digital era, both the technical possibilities and the introduction of new behaviors have participated in the production of large databases storing tremendous amounts of varied information. Recent hot topics such as Big Data, Complex Systems and Network Analysis have been stimulated by this new access to information. One particular topic of interest is the study of how crowds are involved in massive generation of content, whether it be on Wikipedia, Twitter, Facebook, YouTube, or even through the publication of ever growing number of scientific publications. If these datasets are a stimulating opportunity, they are also a challenge. While many research has been done on these topics, we feel there is no simple, generic method to explore this decentralized creation of content, and in particular its dynamic. The platform we propose is generic enough to take input from many kinds of sources, such as scientific publications, online social networks, and many others. The platform is developed with internet based tools only, and could therefore be adapted to provide a user-friendly interface to explore a large dataset of content creation available on the internet.

1.1 Related Works

Several visualizations have been proposed to understand complex systems and large data in general. We introduce the most closely related to our proposition.

ThemeRiver [9] is probably the most famous of these. It allows to represents the dynamic of topics in large collections of documents.

History flows [16]also focuses on dynamic aspects. It is a tool to visualize cooperation and conflict between authors in the process of collaboration, in particular on the web.

In the work by Rosvall et al.[12], alluvial diagrams are used to represent the evolution of communities in networks, and is applied in particular for the visualization of the evolution of research topics in science.

On a more static perspective, numerous tools, frameworks and softwares have been proposed to represent networks in the best possible way. We can cite some of them, among the most famous ones: Gephi[2], Cytoscape[13], Tulip [1].

Several works have also been done on the visualization of dynamic networks; we can cite [3] as a reference on the domain.

The tools we have cited above are either specialized on the visualization of longitudinal aspects, but without information on the internal structure, or, on the contrary, represent this internal structure (network visualization), but only with a static point of view. Our platform is designed to encompass both aspects.

2. MASS COOPERATION DATASETS

In order to illustrate the possibilities and possible practical applications of the tools presented in this paper, we applied them to three large datasets from different fields. In this section, we will present briefly these datasets, and the type of data we extract from them.

For a dataset to be visualized using our platform, it needs to be composed of several productions, that we call Cooperative Productions (CPs). It can be a video, an article, a website, a message, or any other item which can make a reference and be referenced. These CPs are defined by the following properties:

- Name
- Time of publication
- Category (a chain of character, can be omitted)
- List of references it makes to other CPs

Additionally, we need to group these CPs in Cooperation processes. A cooperation process is a set of CPs corresponding to a same topic, a same goal, or any other way of grouping them relevant to the studied dataset. In the following sections, we will detail these properties in 3 example datasets.

2.1 NicoNico

NicoNico, or Nico Nico Douga, is a Japanese video-sharing platform, with functionalities similar to those of YouTube. With officially more than 20 Million registered users, and being ranked among the top 15 most visited websites of Japan, it is a major Web 2.0 platform. It is especially famous for the important community of people cooperating in the creation of complex Music Videos centered on the character of Hatsune Miku. Starting from an original song, many people create videos based on it, with innovation such as dancing, singing, creating new graphics, etc. More information about this character and phenomenon can be found in [8, 10, 6].

We use the dataset described in [7] which covers all 2,622,495 videos published on the network between January 2007 and December 2012.

Definition of a cooperation process
In NicoNico, tags are associated with videos. We automatically detect tags corresponding to songs with more than 500 related videos. These videos compose the cooperation processes.

Definition of a CP

- Name : Name of the Video
- Time : Upload time
- Category : extracted from keywords, examples are: Dancing, Singing, 3D, Animation...
- References: authors include references to other videos in their comments.

Statistics
We obtain 165 cooperation processes, composed by 500 to 7654 videos, with an average of 865 videos.

2.2 Twitter

Twitter is one of the most famous and largest Online Social Networks. In this paper, we consider the diffusion of a particular tweet as our cooperation processes. We used a dataset covering the period between March 5, 2011 and March 24, and which covers most tweets published in Japan during this period. Authors of this dataset claim to have validated that 80% to 90% of all published tweets appear in their dataset. For more information, please refer to [15].

Definition of a cooperation Process
We first counted for each tweet in our dataset the number of time they were retweeted, following the method described in [4]. For all tweets retweeted more than 500 times, we collect all the involved tweets and their information. Each of these sets of tweet form a cooperation flow.

Definition of a CP

- Name : Retweeter's name
- Time : Time of the Retweet
- Category : Distance in the follower network between original author and retweeter
- References: a retweet

Statistics
45 cooperation processes corresponding to retweet chains are detected, involving between 500 and 2100 tweets, with an average of 755 tweets.

2.3 DBLP

Massive cooperation predates the apparition of the World Wide Web. Thousands of researchers around the world cooperate to improve the global scientific knowledge. We use as a dataset the DBLP database [11], and in particular the version including links between papers, as described in [14]. This database is composed of 2,084,055 articles linked by 2,244,018 citations.

Definition of a cooperation Process
As we lack topic information, we define a cooperation process for each article, with all other papers making a direct reference to it composing the cooperation processes. This definition is not perfect, but, as we know that seminal papers tend to act as "flags", that must be cited by everyone working on a specific topic, looking at all papers citing a seminal one is an approximation of a group of works in the same topic. We filtered out all cooperation processes with less than 500 elements.

Definition of a CP

- Name : Publication Title
- Time : Date of Publication
- Category : Venue of publication
- References: a citation to another paper

Statistics
After filtering, we obtained 41 citation flows, composed of between 500 and 3651 papers, with an average of 664 papers.

3. DESCRIPTION OF THE PLATFORM

The platform we propose is composed of two parts: the time series visualization and the cooperation flow visualization. The time series provides a global understanding of the different cooperation processes studied, together with global

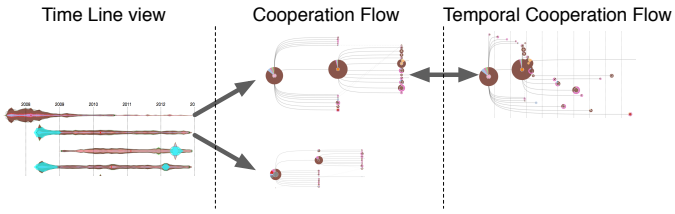


Figure 1: Schema of the possible navigation between the displays of Cooperation Explorer

indicators on them. In this view, only the global properties are represented, not the individual agents and their interactions. From this global view, it is then possible to select any cooperation process and to visualize its inner details in the cooperation flow view.

In this second tool, in which the details of the cooperation is displayed, several options are possible such as positioning according to time or to step of cooperation, selecting the number of nodes displayed, etc. The navigation between these different displays in represented in Fig. 1

3.1 Temporal trends

When we are interested in a cooperation process, it is often useful to have first a global vision of it. We would like to be able to answer general questions such as: when did this process started? Is it already finished? Is it becoming more or less popular? Are there some patterns in its popularity? These are the global properties of this particular cooperation.

3.1.1 Macro-level visualization: time series

The visualization we propose excludes the role of each specific element, to represents the process as a whole. To do so, we choose to transform our data in time series, as much work exists on the topic of time series analysis. For a given dataset, we define a time step, which can be any period of time (minute, day, year, etc.) and count the number of CPs published for each category in each time step. For each Cooperation Flow, we obtain as many time series as there are categories. We display them as a shape, as shown in Fig. 7. The shape is constructed as a cumulative area chart augmented with a mirror image of itself, to have a symmetric shape. The lecture of it is identical to a normal cumulative area chart. We choose this shape instead of a normal cumulative area chart because we want to represent several of these shapes on a same plot with a single time axis. Therefore, the shape is not framed by the axis, and when displayed on top of each other, it becomes more natural to have a horizontally symmetrical shape, as represented on Fig. 8. A similar observation has been done by the authors of ThemeRiver [9].

By displaying several shapes on the same chart, we are able to visually compare them. Examples of interesting observable facts include (but are not limited to):

- The relative importance of different categories along time
- The presence of bursts at a particular location, or following a fix period
- Differences between cooperation processes starting at different times

We complete this tool with some metrics:

3.1.2 metrics and graphics

Lifespan

For each cooperative Process, we compute its lifespan, defined as the time between the first not null value of the time series to the last occurrence of 3 consecutive not null values. This limit is arbitrary, but the objective is to give an end to a time series, potentially infinite, as a new CPs can always occurs in the future. If these 3 non-null values are the last 3 values of the time series, we consider the cooperation process as "still alive". The distribution of the lifespans is displayed as a bar chart.

Normalized centroid

We compute the normalized centroid of each cooperative flow. The centroid of the time series is the step such as there is as many CPs before and after it. We normalize it by computing:

$$NormalizedCentroid = \frac{centroidTime - birthTime}{deathTime - birthTime}.$$

A normalized centroid inferior or superior to 0.5 reflect the fact that most of the CPs where produced in the beginning or in the end of the lifespan of the cooperation process. The distribution of the normalized centroid is displayed as a bar chart.

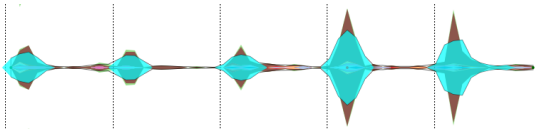


Figure 2: Visualization of periodic bursts in a temporal trend. Detected bursts appear in translucent blue color. In this case, we can observe yearly events.

Burst detection

Burst detection is a common problem on time series. A burst is defined as a period of time during which the time series reach temporarily exceptionally high values. In a cooperation flow, such a burst can typically appears in the beginning (initial burst), at a given moment, driven by internal events (new popular CP), or external factors. An interesting case is when this external factor is not unique but periodic, typically daily or yearly events. We therefore implemented a research of such periodic bursts. We implemented the burst detection with a simple but effective technique, presented in [17]. We represent the bursting period with a translucent color as seen in 2. We compute normalize burst positions in a similar manner as we computed normalized centroid, and the summary of the most common burst positions detected is also represented as a bar chart.

We found 5 cooperation processes with periodic bursts in the NicoNico dataset, and we checked that all of them cor-

responded to yearly events (songs about Christmas, Halloween, etc.).

3.2 Micro-level visualization

Whereas the time series visualization allow us to have a quick understanding of global properties, it is often useful to have more insights in the details of what is happening inside each cooperative topic. In this second display, we combine a visualization called cooperation flow together with some alternatives displays and indicators, each of them emphasizing one aspect of the studied cooperative topic.

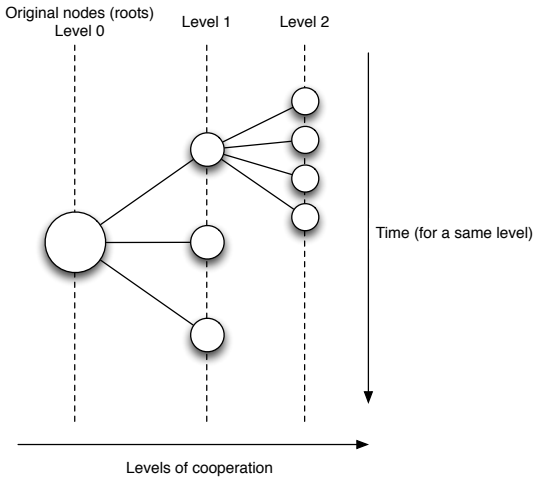


Figure 3: Mechanism of the cooperation flow visualization.

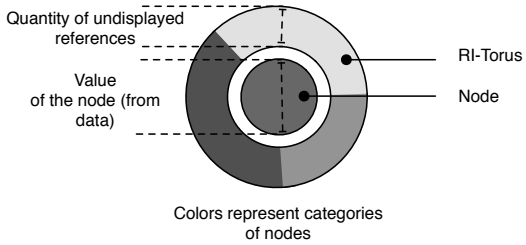


Figure 4: Schema of the representation of a node

3.2.1 *Cooperation Flow*

To represent the details of the process of cooperation, we use a type of visualization described in [5]. This visualization, called Cooperation flow, allows us to represent in a single visualization the key points of the details of the process. Its mechanism is represented in Fig.3. The idea is that, through the interface, we specify the maximum number of nodes that we want to display, n. An algorithm compute which are the n most important elements for the cooperation in the current process. These nodes are then displayed, together with their relations, as a network organized by steps of cooperation. More formally, the step of a node is defined as the length of the shortest path between this node and a root, that is to say a node without any reference to other nodes. The nodes which are not considered important enough to be displayed are, however, not simply omitted. By using a feature called Reuse Indicator Torus (RI-Torus), a summary of these nodes appears around their last displayed ancestor, as summarized in Fig. 4.

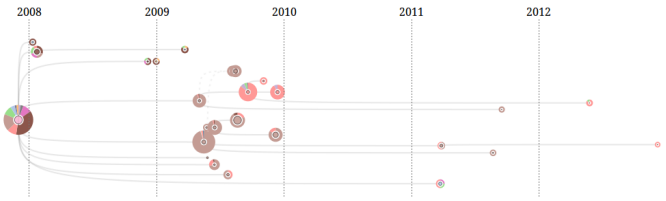


Figure 5: Example of a cooperation flow where the x position represents time

3.2.2 *Temporal Cooperation Flow*

One interesting property of this visualization is that nodes situated on overlapping y values are necessarily ordered in a chronological order from left to right. Therefore, it is possible to switch to a temporal representation without changing the y position of nodes. This is illustrated on figure 5.

3.2.3 *Complementary visualizations and metrics*

We added to this visualization a set of informative visualization and metrics, each of them focusing on a specific aspect of the cooperation. These tools are based on the same data as the cooperation flow visualization. All of these tools are not affected by the selection of nodes we make for the flow visualization, they are based on all available information.

Impact of main CPs

We observed that one characteristic which can vary greatly between cooperation flows is the importance taken by the most important productions. In some cases, a single production, or a small subset of them, can generate most of the CPs, that is, most of the CPs will directly reference it as a unique source, either during the whole lifetime of the flow, or just during a given period. To study this, we propose a visualization in stacked area of the impact along time of the top 5 nodes, toped by the impact of all remaining nodes (Fig. 6). The lifetime of the flow is split in 10 sections. The impact of a given CP during a given section is computed as the number of CPs published during this period that reference it. We use a black and white scale to avoid confusion with the categories of CPs, already represented by colors.

Together with this visualization, we propose a metric to measure this effect, called CSC, for Cooperation Source Concentration.

$$CSC = \frac{\sum_{v \in Top1} |\{u : (u,v) \in E\}|}{|V| - 1}$$

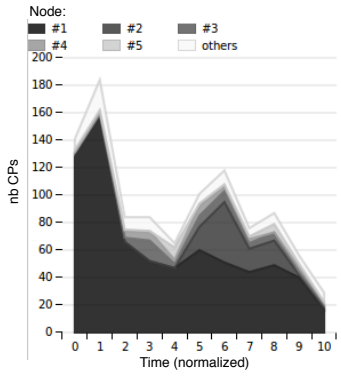
Where $Top1$ is the set of the 1% nodes with the highest in-degree. This metric vary between lim_0 and 1, where lim_0 is the case where all nodes have the same in-degree, and 1 is reached when all nodes are successors of a single original source. (star-like network) We give the average values of CSC for our 3 datasets in Table 1. We can observe large differences, with NicoNico having the strongest CSC and Twitter the lowest.

Sustainability of the cooperation

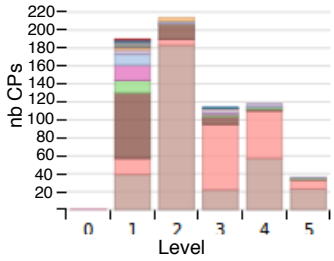
We observed that in some cooperation flow, there is not much cooperation after the first few levels -the number of CP by level follows a fast shrinking trend- while, in others, it is not the case. This might reflect the ability to renew the

	NicoNico	Twitter	DBLP
Average CSC	0.92	0.21	0.65

Table 1: Average value of CSC by dataset. CSC represents how important is the role of the top 1% users in the cooperation.



(a) Impact of the main CPs along time, in a particular cooperation flow. In this example, we can see that 2 or 3 nodes are the source of most cooperation. For example, the increasing number of videos in period 6 is mainly due to the popularity of a single node.



(b) Sustainability of the cooperation. In this case, we can observe a shift in the categories of CPs as with progress in the levels of cooperation. This pattern is common on NicoNico.

Figure 6: Additional analysis tools

interest in the trend by new CPs. We propose a visualization of this effect by a stacked bar chart graph (Fig. 6). Each bar represent a level, and we simply count the number of videos of each type published in each level. Together with the general trend, this chart allows to see a change in the categories correlated with the level. The color used for the categories are coherent with the ones used in the cooperation flow chart.

The indicator we propose to summarize this chart is SC, Sustainability of Cooperation. It is defined as the average of the variations of the number of CPs between successive levels, pondered by the number of CPs in the first of the two:

$$SC = \frac{\sum_{i=1}^{nl-1} \frac{nbCP(i+1)}{nbCP(i)} * (nbCP(i+1) + nbCP(i))}{nbCP(1) + 2\sum_{i=2}^{nl-2} nbCP(i) + nbCP(nl-1)}$$

with $nbCP(i)$ the number of CPs at level i, and nl the number of levels. SC=0 if there is no production after the first level. SC >1 if the number of CPs tends to grow with each level. The lower the SC value, the less CPs tend to generate new cooperation. In Table 2, we represent the average value of SC for our datasets.

	NicoNico	Twitter	DBLP
Average SC	0.23	0.39	0.44

Table 2: Average value of SC by dataset. SC represents the average ratio between the number of videos published at step i and $i+1$.

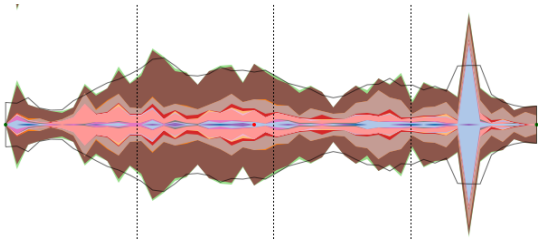


Figure 7: Visualization of the temporal trend of a cooperation flow

4. EXAMPLE VISUALIZATIONS

In this section, we briefly present some examples to show the interest of our visualization.

4.1 Temporal trends

Our visualization allows to display on a same timeline the time series of several temporal trends. We can therefore compare them, observe some typical behaviors or spot outliers. In Fig. 8, we show an example of this view on our example dataset. We can observe very different properties. For instance, in Twitter, we see a typical bursting behavior, followed by a rapid decay. Most of the productions, i.e., retweets, occur in the beginning. On Nico Nico, the trends are more long lasting, bursts are not as important. People continue to publish videos at the same rate for years. Finally, in the citation dataset, we observed more varied patterns, and even some "increasing" trends, for which the number of papers published increase from years to years.

4.2 Cooperation flows

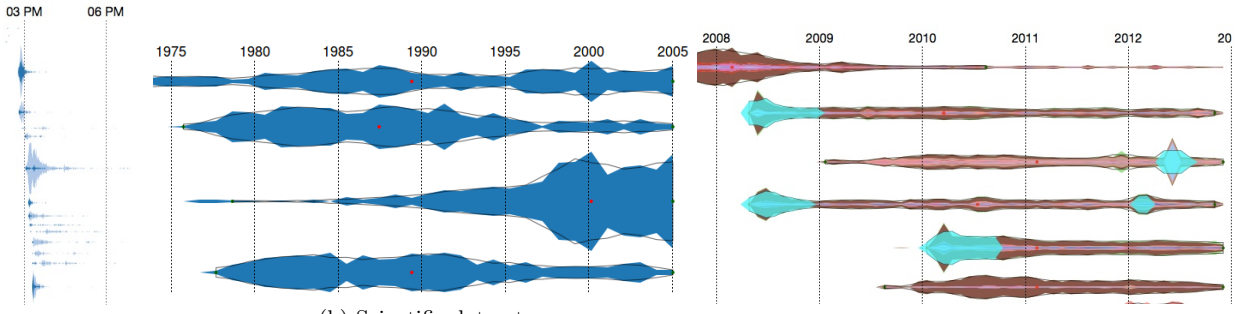
4.2.1 Deep study of one dataset: NicoNico

NicoNico is the richest and the most complex of our datasets. In fig. 9, we show 2 typical flow from this network. We can make the following observations, also valid on most other flows:

1. There is only one original source, and most of the cooperation is made directly from this source, as we can judge by the large RI-Torus
2. Most important nodes for the collaboration are on the first level, they directly reference the original node only
3. The cooperation is more wide than deep, there is not much cooperation at a level greater than 3.
4. Although many categories (colors) are present, each node seems to generate a specialized cooperation: RI-Torus are mostly of a single color, not always the same.
5. There is no strong correlation between the number of view of a video (area of inner circle) and its capacity to generate cooperative behavior (torus area)

4.2.2 Comparison of datasets

In fig. 10, we present two visualizations typical of the other datasets. We can immediately spot some differences. In the tweet dataset, cooperation is deeper, and we tend to see the formation of chains, long but without many bifurcations. More important nodes are not necessarily situated



(a) Twitter dataset

(b) Scientific dataset

(c) Nico Nico Dataset

Figure 8: Examples of temporal trends

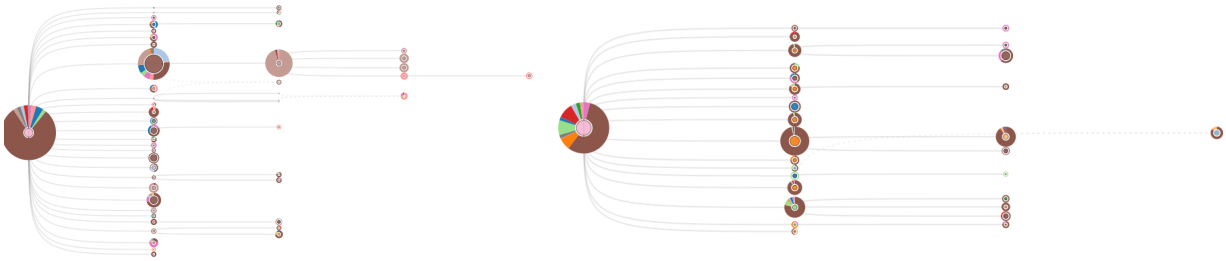
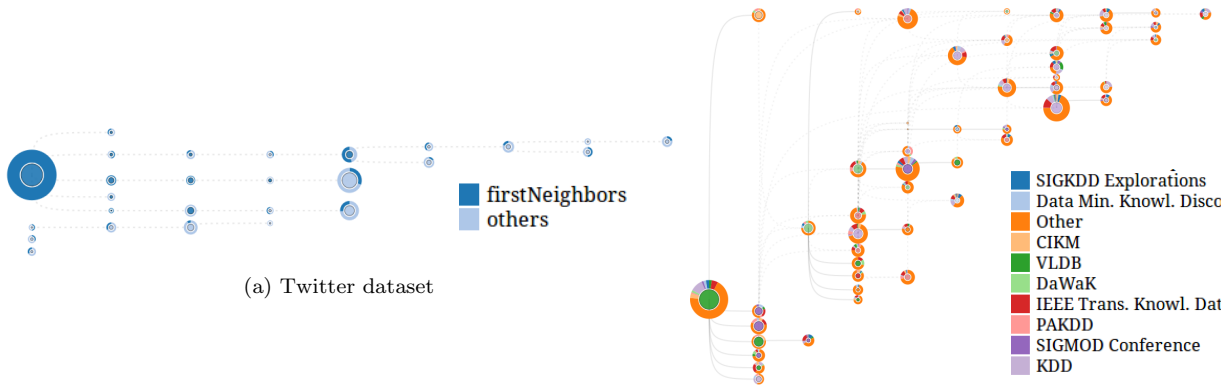


Figure 9: Examples of typical cooperation flow in NicoNico



(a) Twitter dataset

(b) DBLP dataset

Figure 10: Examples of typical cooperation flow in Twitter and DBLP

at the first step, but can occur deeper. There seems to be a stronger relation between the popularity of the node and its role in the cooperation. There is not a single source.

In the citation dataset, we immediately spot a large number of nodes making references to several others. These nodes with many references are important in the cooperation. Nodes at a deep level seem to generate as much cooperation as those in the first levels. There also seems to be a lesser concentration in the cooperation generation: a larger fraction of nodes are referenced by other important nodes, and the gap is less important between the top influential nodes and the ordinary ones. Exploring in further details the properties of the different datasets is beyond the scope of this paper.

5. CONCLUSION

In this paper, we have presented a platform to explore mass cooperation, and a set of tools to explore different aspects of this type of cooperation. Our conception of such visualization was driven by our previous experiences in the exploration of large datasets formed by cooperation, and the difficulties encountered to understand the underlying mechanisms.

We also presented some complementary visualizations and metrics that focus on several aspects of the data, with different granularities, and can also help to apprehend it.

In the future, we hope that other researchers will use this platform and help to improve it, either by their remarks or extending the possibilities. In this prospect, we release its source code, altogether with an interactive online version, so as interested researchers could work with it as easily as possible. In particular, it could be interesting to add metrics and statistics, such as a one could choose the more interesting indicators in his case. The source code and browsable example is available on the website of the first author.

Another future possibility is to propose Internet applications based on this visualization to the destination of final end users. For example, one can think of a plug-in for Google Scholar allowing to browse research topics.

6. ACKNOWLEDGMENTS

We thank Fujio Toriumi for collecting the Twitter dataset, and allowing us to make use of it in this work.

7. REFERENCES

[1] D. Auber. Tulip, a huge graph visualization framework. In *Graph Drawing Software*, pages 105–126. Springer, 2004.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *ICWSM*, pages 361–362, 2009.

[3] S. Bender-deMoll and D. A. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7(2):1–38, 2006.

[4] R. Cazabet, N. Pervin, F. Toriumi, and H. Takeda. Information diffusion on twitter: everyone has its chance, but all chances are not equal. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*, pages 483–490. IEEE, 2013.

[5] R. Cazabet and H. Takeda. Understanding mass cooperation through visualization. *ACM Conference on Hypertex and Social Media*, 2014.

[6] R. Cazabet, H. Takeda, M. Hamasaki, and F. Amblard. Using dynamic community detection to identify trends in user-generated content. *Social Network Analysis and Mining*, 2(4):361–371, 2012.

[7] M. Hamasaki and M. Goto. Songrium: a music browsing assistance service based on visualization of massive open collaboration within music content creation community. In *Proceedings of the 9th International Symposium on Open Collaboration*, page 4. ACM, 2013.

[8] M. Hamasaki, H. Takeda, and T. Nishimura. Network analysis of massively collaborative creation of multimedia contents: case study of hatsune miku videos on nico nico douga. In *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, pages 165–168. ACM, 2008.

[9] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, 2002.

[10] H. Kenmochi. Vocaloid and hatsune miku phenomenon in japan. *Proc. of InterSinging 2010*, pages 1–4, 2010.

[11] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer, 2002.

[12] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[13] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[15] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1025–1028. International World Wide Web Conferences Steering Committee, 2013.

[16] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM, 2004.

[17] Y. Zhu and D. Shasha. Efficient elastic burst detection in data streams. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 336–345. ACM, 2003.