Temporal Analysis of the Wikigraph

Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, Stefano Millozzi

Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, 00198 Roma, Italy
Email: {*buriol, castillo, donato, leon, millozzi*}*@dis.uniroma1.it*

***Abstract*— Wikipedia (www.wikipedia.org) is an online encyclopedia, available in more than 100 languages and comprising over 1 million articles in its English version. If we consider each Wikipedia article as a node and each hyperlink between articles as an arc we have a "Wikigraph", a graph that represents the link structure of Wikipedia.**

The Wikigraph differs from other Web graphs studied in the literature by the fact that there are timestamps associated with each node. The timestamps indicate the creation and update dates of each page, and this allows us to do a detailed analysis of the Wikipedia evolution over time.

In the first part of this study we characterize this evolution in terms of users, editions and articles; in the second part, we depict the temporal evolution of several topological properties of the Wikigraph. The insights obtained from the Wikigraphs can be applied to large Web graphs from which the temporal data is usually not available.

I. Introduction

In the past decade, the Web has experienced a very fast growth rate: recent estimates [1] indicate that the indexable Web exceeds 11.5 billion pages. The Web is also very dynamic: pages are modified, created and deleted continuously. Because of this huge amount of changing data, search engines have to constantly afford the burdensome task of updating their index in order to keep an up-to-date copy of the current Web.

The study of the evolving Web has been mainly focused on the degree and the frequency of changes in the Web pages. The statistics collected on sequential crawls of the Web have been used to develop incremental crawling algorithms to increase the average "freshness" of the pages in their indexes, and are being used to develop time-aware ranking algorithms able to provide the final user with the most relevant results available.

Ntoulas and Cho [2] observed that the evolution of the hyperlinked structure is more dynamic than the evolution of the textual contents of the pages. After one year, the percentage of links still present in the Web is only 24% against a number of unchanged pages that reaches 50%. This fast pace of change in the Web graph is very important if we consider that the hyperlinked structure is the basis of many algorithms that assign an authoritativeness score to the pages.

It is worth to underline that extracting the link structure of the Web at a specific point in time is not possible. Downloading pages from the Web, in a certain way, resembles watching the sky on a clear night: what we see reflects the state of the stars at different times, as the light had to travel different distances. What we obtain by Web crawling is not a "snapshot" of the Web, because it does not represents the Web at any given instant of time [3].

An attempt of capturing the dynamics of the Web can be made by collecting a series of frequent static snapshots by sequential crawls. From the study of these snapshots, it can be inferred if a page has been modified or deleted during a certain time frame. However, it is not possible to determinate exactly the instant when the update or deletion occurred. Also, relying on the update time provided by the HTTP server responses is not correct [4]: only 40% of the HTTP headers present time information (i.e. creation and last update time).

Kumar et al. [5] overcame this problem considering the evolution of the **Blogspace**, this is, the set of Web logs (or *blogs*). A Blog is commonly a page that contains a series of dated entries (or *posts*) ordered from newest to oldest. Each time that a new entry is inserted, the page can be considered updated. In this way, all the information concerning the "time" can be directly extracted from the date of each individual entry.

Our approach to the study of the evolution of Web graphs over time is to study the Wikipedia (`http://www.wikipedia.org/`), an on-line and free content encyclopedia written in more than 100 languages and with over 1 million articles in its English version. Details about the evolution of the Wikipedia on time are available for every article and every version of an article in the Wikipedia.

We present a study of the hyperlinked graph originated from the link structure of the articles in this Encyclopedia. This work aims at verifying if any evolving trend is observable in the statistical properties of the articles and/or the Wikigraph, and how different measures evolve together over time. We want to stress that, up to now, very little research work has been devoted to the evolution of the statistical and topological properties of hyperlinked graph as Web graphs, Blog graphs and Wikigraphs.

Besides the fact that the Wikipedia has grown over time, this study reveals interesting properties of the Wikigraph:

- the Wikipedia in general has become "denser" over time both in terms of contents and hyperlinks,
- the degree of vandalism seem to have increased but still a low fraction of the updates are related to correcting vandalism, and this is usually corrected very rapidly,
- in some aspects the Wikigraphs appears quite mature, while in other aspects it is still evolving rapidly.

In many aspects the Web graph and the Wikigraphs are very similar, and to a certain extent this study provide several hints about how the Web at large is evolving and may continue to evolve in the future, in particular with respect to its connectivity. This article aims at gaining insights about the evolution of an hyperlink structure from the study of a coherent, information-rich corpus.

The next section discusses related work on this topic. Section III presents our data set and section IV depicts the growth of the Wikipedia in the studied period. Section V studies the dynamics of the article updates. Sections VI and VII the evolution of the hyperlink graph at the microscopic and macroscopic level respectively. The last section presents our conclusions.

II. Related Work

Web graph: The study of the topological properties of the Web graph started in [6], [7]. A more complete analysis of the Web graph was later presented in [8] where many measures of the Web were presented together with the bow-tie picture, a macroscopic characterization of the Web structure. Later, the bow-tie structure was refined by an extensive study of a large sample of the Web provided by the WebBase project [9].

Temporal evolution: Cho et al. [10] presents the results of an experiment conducted over 4 months. The authors daily crawled 270 sites in order to measure the rate of change and the lifespan of each page. A Poisson process was used to model the rate of change and compare the efficiency of different crawling strategies. The authors also described the architecture of a incremental crawler able to keep up the index with the evolving Web.

Other studies of the temporal evolution of Web graphs include [4], [11], [12], [2]. Most of them traced a set of pages in order to compile some statistics about the frequency and rate of the changed pages and the percentage of pages that are deleted or created every year. The search engine perspective is dominant in all of them.

For instance, Fetterly et al. [12] expanded the work of [10] both in terms of coverage and sensitivity to changes. They found out that good predictors of future changes in the Web are the top-level domain pages, and relate document size and history to the freshness of a Web page collection.

A search engine-centric approach is followed also by Ntoulas et al. [2]. The authors crawled 154 “popular” sites for a year and revealed a high dynamical behavior of the Web. But, despite of the high rate of newly created pages, the ‘new contents’ introduced are less than 5% of all changes introduced. They also observed that the Web link structure is even more dynamic with more than 75% of new links every year. Moreover they found out that, for pages with significant changes over the time, the degree of changes tends to be highly predictable and observed that this results can be used to crawl proper portions of the Web.

To capture the relation between the popularity, authority and time, a few recent studies [13], [5] have presented observations that directly couple hyperlinks with temporal data.

Models for analyzing the evolution of the Web graph were presented in [13], [5]. In particular Kraft et al. [13] defined the notion of *TimeLinks* and extract some statistics over the data. Kumar et al. [5] introduced the notion of *time graph* and conducted a series of experiments in order to trace the formation and the development of communities in the Blogspace and to detect burst of activity within them.

Wikipedia: On the Wikipedia, our previous paper [14] studies the bow-tie structure in the non-English Wikipedias, and Reference [15] deals with the graphical representation of the history of an article.

The Wikipedia is also an excellent source of data for other Information Retrieval tasks. For instance, currently the INEX initiative `http://inex.is.informatik.uni-duisburg.de/2006/` provides a dataset of Wikipedia articles annotated with topic and relevance assessments for research purposes.

As for the quality of the content, [16] compares Wikipedia to Encyclopedia Britannica and [17] tests quality metrics that can be derived from automatically-extracted features of Wikipedia articles. Finally, Wikipedia itself provides some information about itself on `http://en.wikipedia.org/wiki/Wikipedia:Statistics`.

III. Data set

Wikipedia is an on-line and free content encyclopedia. The first few English pages were published in January 15, 2001. Four and half years later, Wikipedia has more then 1 million articles.

There are a number of reasons that lead us to consider this encyclopedia a good dataset for Web graph-type experiments:

- *Diversity*: the encyclopedia collects pages written by a number of independent and heterogeneous individuals. Each of them autonomously decides the content of their articles with the only constraint of a prefixed layout. The autonomy is a common feature of the content creation in the Web. The Wikipedia authors’ community is comprised by members that are pushed by the only wish to make available to the world concepts and topics that they consider meaningful.
- *Metadata*: Wikipedia provides time information associated with nodes. Moreover, it provides old information: time information regarding the creation and the updates for each page on the dataset.
- *Independence of external links*: Wikipedia articles link mainly to articles on the same dataset.

All the data from the Wikipedia is freely available at `http://download.wikimedia.org/`. The dataset is a large, compressed XML file containing information about **pages**, and inside each page, information about all the **revisions** the page has undergo.

For each revision, the time stamp, author, editorial comment and full text of the version of the page is available. The **author** corresponds to the name of a registered user in the case of normal edits, to the IP address of the originating computer in the case of an anonymous edit, and to the program name in the case of a (semi-)automatic edit. Automatic edits are done to adapt articles to newer standards and automatically create links, and except in the case of very trivial syntactic changes, are usually reviewed by a human operator.

We consider articles only in the main name space, this is, encyclopedic articles themselves, as opposed to templates, discussions, user pages and other administrative pages.

Sometime a page is a **redirect**, which is an alias pointing to another page (e.g.: “Einstein” points to “Albert Einstein”). Redirect pages do not have content on their own, and are used mostly to provide several different access points to the same article. There are many redirects in the Wikipedia and about 52% of the pages in the main namespace are in this category.

In the following we refer to a page in the main name space that is not a redirect as an “**article**” or simply a “page” when the context is clear.

In order to generate a graph from the link structure of a dataset, each article corresponds to a node and each hyperlink between existing articles to an edge. An article also might contain some external links that point to pages outside the dataset, but they are usually only a few. These kind of links are not considered for generating Wikigraphs, since we want to restrict the graph to pages into the set being analyzed.

Also, when a page u is a redirect (alias) pointing to a different article v, we removed the page u from the graph and re-wire all of its in-links to point to the destination page v. Interestingly, we noticed that there are a few redirect loops in Wikipedia, this is, articles redirecting the user to other article that redirects the user back to the original one. In that case, we took the older article in the redirect loop as the representative of the loop.

For analyzing the evolution of the Wikipedia, we generated 17 **snapshots** of the Web graph at different points of time. We started with the Wikipedia graph as of January 1st, 2002 (14,247 articles) and generate one snapshot every 3 months until April 1st, 2006 (1,064,757 articles). We used the COSIN library for the analysis [18], this is a software for processing large graphs that implements several algorithms in semi-external memory.

IV. The Growth of Wikipedia

The number of articles, shown in Figure 1 (a) presents a remarkable growth, roughly doubling itself in size every year for the last three years, this corresponds to a 6.2% montly growth. In the same period, the number of updates (b) has been growing at a rate of 11.3% per month. This means that in the most recent Wikipedia “versions” each individual article receives more attention that in the previous ones.

The number of different visitors and different editors have also been growing at a fast peace, as shown in Figure 1 (c). The number of unique visitors has grown at a rate of roughly 15% during the last three years, while the number of unique editors (d) has grown at a rate of roughly 13%.

Interestingly, not only the number of articles has grown, but also individual articles have become longer. In [15] it is shown that pages with more than 100 edits grow steadily over time. Actually, if we consider **all** the pages that existed three years ago (January 2003) and plot their size, as in Figure 2, we observe that they become longer, growing linearly during the last two years at a rate of 1 KB of text every 6-8 months.

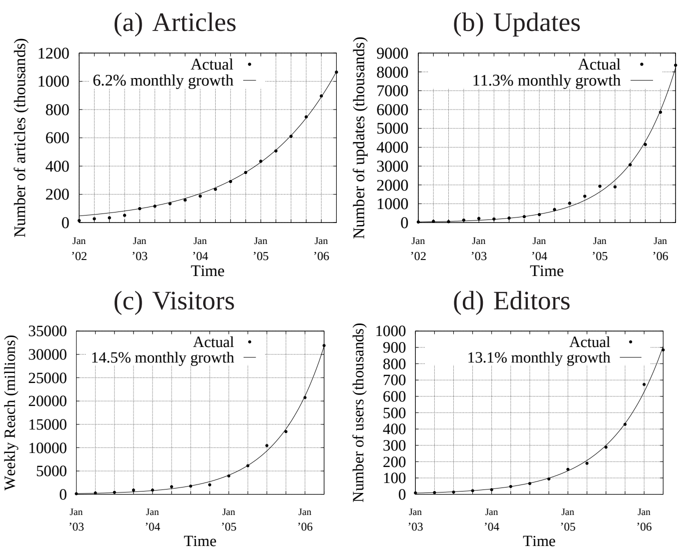

Fig. 1. Growth of the Wikipedia: (a) number of articles, (b) number of updates, (c) unique visitors, in terms of monthly average provided by Alexa and (d) distinct editors.

Newer articles follow a similar growth process; we are also including in the plot the articles that were created during 2003, 2004 and 2005. The growth rate of newer articles does not seem to be accelerating and in fact appears to be lower than the growth rate of older articles, so older articles attract more edits than newer ones. This may be interpreted as a sign of maturity of the coverage of Wikipedia, but not of the depth of particular articles (many of them are marked as “stubs” and tend to be very terse), so we expect that the average length of the articles continues to grow for a few more years.

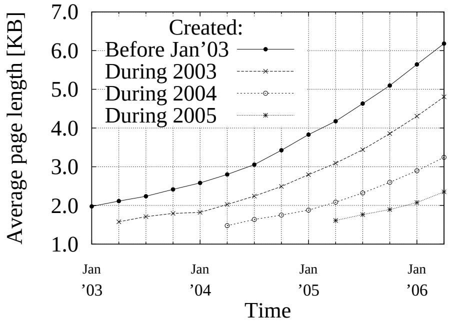

Fig. 2. Growth of the size of articles written on or before a certain date.

Even if the number of edits is increasing, the average change of a page that is edited, measured in terms of how many bytes the page gains or looses, has remained roughly the same during the entire Wikipedia history (except at the very beginning). On average, on each edit a page changes from 300 to 500 bytes, this is roughly equivalent to a short paragraph of text.

V. The Dynamics of the Updates

The updates of pages are a stream of data in the form (*time stamp*,*article*,*user*) representing the events of creating new revisions of pages. The distribution of updates per page follows a power-law, as shown in Figure 3 (left). About 53% of the pages have received more than 10 updates, and about 5% more than 100 updates.

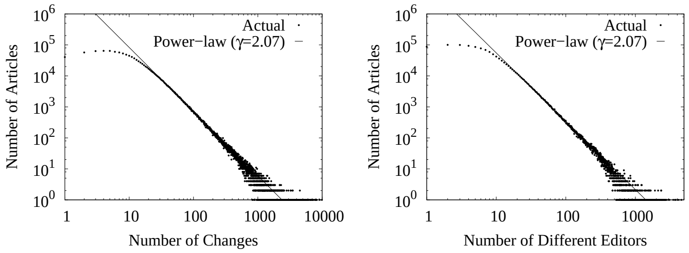

Fig. 3. Distribution of the updates per article. Left: number of changes, right: number of different users involved on each article.

The reason for this skewed distribution appears to be twofold: first, the updates might be related to page views, in the sense that highly-visited articles are updated more often, so the distribution of updates follows the distribution of page importance. Second, the Wikipedia is built as a community and there is a prominent link to see the last 50 changes plus the option to "watch" an article, to receive alerts when it is changed. This may cause that edits made to an article attract even more edits, as explained later in this section.

The distribution of the number of different users involved on the edition of each article is also very skewed, as shown in Figure 3 (right). Having a single editor is rare (about 7.5% of the articles), but some articles may have a large number of editors: about 50% of the articles have more than 7 different persons involved and about 5% of the articles have more than 50 different editors. The average number of updates per user has dropped by about 30% in the last two years, as shown in Figure 4 (right).

In any case, the fraction of articles that are updated is very high. Over 80% of the articles are updated in the three-months period we are considering for these snapshots. This fraction has remained basically constant during the last two years, as shown in Figure 4 (right).

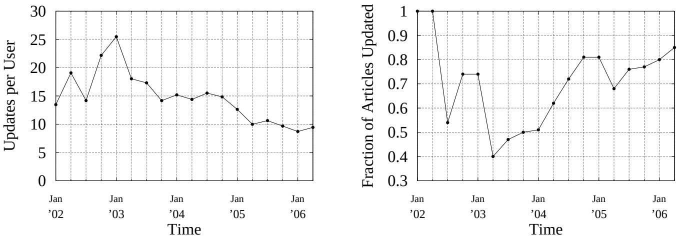

Fig. 4. Left: average number of edits per user. Right: fraction of articles existing in one period that are updated by the end of the same period.

There are many interactions among the actions of different editors. In general after one article is edited, there is a 7% chance that the same article gets edited again *by a different user* during the next hour. This probability raises to 13% if we consider 6 hours and 22% if we consider a 24 hours-period. This has remained constant during the history of Wikipedia. These numbers reflect actual collaboration among users and do not include the number of "reverts".

Reverts can be done by any registered user, and with this one-click operation an article can be taken back to a previous version. This is done mostly to fight vandalism. We detected when an update is a revert by searching for the string `revert` or `rv` in the comment of the edits (this is inserted automatically).

The fraction of editor actions that are reverts is in general small, but it has been steadily increasing in the last years, as can be seen in Figure 5. This may signal an increasing amount of vandalism per page, although in general the number of reverts per edit is less than 6%, and does not seem to be related to the number of anonymous edits, which has remained consistently between 20% and 30% over the last years.

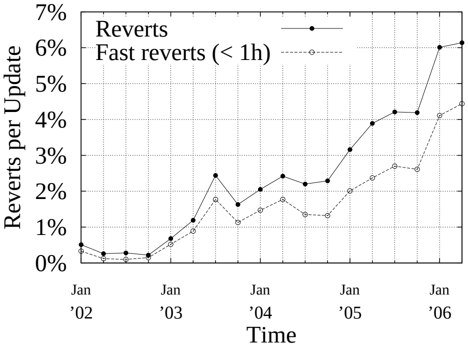

Fig. 5. Percentage of editor actions that are reverts.

As the main reason for reverts is to correct vandalism, it was observed in [15] that in general about half of a certain type of vandalism (the vandalism that involves a mass deletion of content in a page) is corrected within the next three minutes after it is done. Our measurements indicate that the percentage of reverts that are executed in less than one hour has remained constant over time and is around 70%.

Sometimes the system of reverts is also used editorially, when one editors disagrees with other and rejects his/her changes. This is considered rude in the Wikipedia and can be replied with a second revert by the affected user, reiterating his/her editions. This creates a phenomenon known as a "revert war" in the Wikipedia, and can be difficult to resolve. In fact, there are technological tools such as protecting a page, as well as social tools such as arbitration committees that can help settle these edit wars.

In November 2004 a 3-revert rule was established: no user should revert a page more than three times in a 24-hours period. This had an effect in the rate of double-reverts that dropped from 7.8% to less than 4% almost immediatly. The rate of double-reverts has continued under 5% after that, as shown in Figure 6.

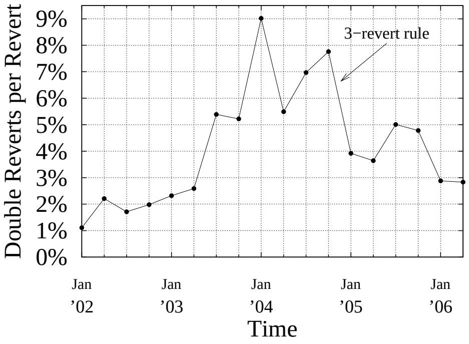

Fig. 6. Fraction of double-reverts per revert.

While the amount of updates is increasing in the Wikipedia, if we focus on individual pages, we can see that this change is not homogeneous. Some entries are updated in response to external events (such as the articles related to candidates in a political election), while other entries have their updates more evenly distributed in time (such as the articles on biology or mathematics).

An attempt to characterize the differences among different articles is to use clustering, as suggested in [19] in the context of information diffusion in Blog posting.

In Figure 7 we have clustered by the k-means algorithm the "update profile" of a set of pages into 4 clusters. The set of pages we clustered is all pages that have existed for at least three years. The obtained clusters can be characterized as: (1) pages that are updated evenly in time, (2) pages that have been more updated than usual in the last 9 months, (3) in the last 6 months and (4) in the last 3 months.

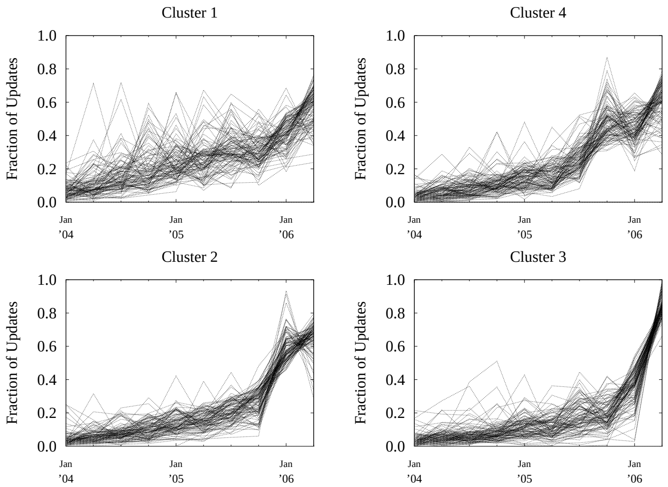

Fig. 7. Clustering of pages per update profile.

By looking at the list of articles on each cluster, there seems to be no topical relationship between them, and no single event explains several changes to the Wikipedia. While an external news event may trigger a large number of updates on a single article, no single event has triggered a massive modification of many pages in the Wikipedia simultaneously.

VI. The Evolution of the Hyperlinks

As expected, the Wikigraph is a scale-free network in which the indegree distribution follows a power-law, i.e. the probability that a node has indegree i is proportional to $\frac{1}{\gamma^i}$, for $\gamma > 1$. We found $\gamma = 2.00$, similar to the value that has been observed in the indegree distribution of Web graphs (2.1) [8], [9]. For the outdegree the distribution appears to be log-normal or double-pareto with an exponent in the tail (articles with more than 30 out-links) of $\gamma = 2.46$. Both distributions are depicted in Figure 8.

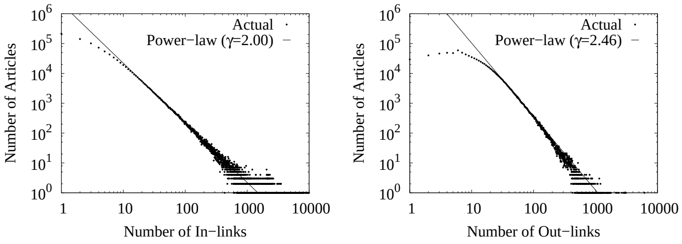

Fig. 8. Degree distribution in the English version of the Wikipedia as of January 1st, 2006. Left: indegree, right: outdegree.

The power-law exponent of the indegree has remained remarkably constant even when the graph has grown substantially in the studied period (from ≈14,000 articles to ≈1,000,000).

The Wikipedia graph is becoming denser. During the last two and a half years, articles have grown from an average of 7 out-links per article to an average of 16 (on average one new reference every 100 days). This is shown in Figure 9.

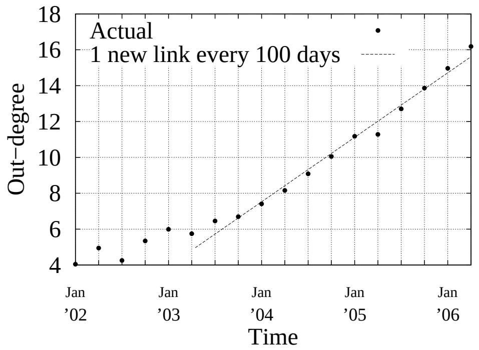

Fig. 9. Average number of links per article over time.

This is a trend observed in other social networks by Leskovec et al. [20]. They observed that for several networks, the number of edges grows exponentially with the number of nodes. The exponent is typically small, in the range 1.1 to 1.6 depending on the particular network.

In the Wikipedia, one possible explanation of this increase in the number of out-links could be that, as the size of the articles is getting longer, there are more concepts that should be linked. However, we can factor out the effect of size and note that in fact, now there is *more* text per each out-link than in the past. From an average of 200-250 bytes of text per out-link in the first three years of the Wikipedia, this quantity has grown to about 300-350 bytes per out-link in the last year, as depicted in Figure 10.

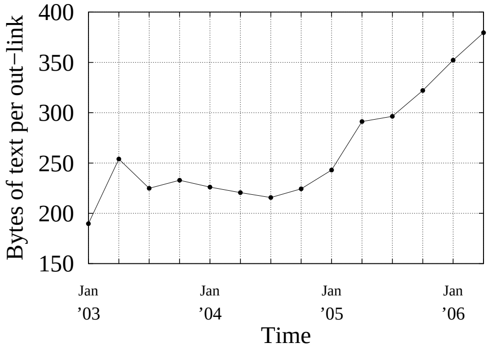

Fig. 10. Average number of bytes per out-link over time.

We also studied the distribution of PageRank [21], a link-based ranking function that assigns to every page in the Web graph a number corresponding to the probability of visiting the page when following links at random (on a modified version of the Web graph in which "sinks" are avoided). PageRank represents mesoscopic (mid-range) properties of the graph, while the in-degree is a microscopic characteristic. The tail of the distribution of PageRank for the usual parameter setting follows a power-law [22]. In our case, we observe a power law distribution with exponent $\gamma = 2.1$. Previous measures for the Web graphs [9], [23] have measured the same exponent.

Usually the correlation between PageRank and indegree is very low in Web graphs. Interestingly, in the Wikigraph this correlation has increased in the last years, basically following the densification of the graph, as shown in Figure 11.

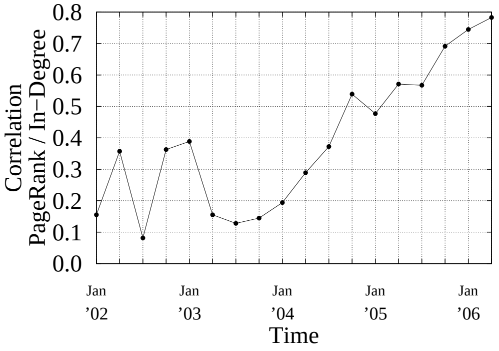

Fig. 11. Correlation between PageRank and indegree over time.

This means that many of the new links that have been appearing point to pages that already have high PageRank. Potentially, this could mean that the low PageRank/indegree correlation observed in Web graph is a transient phenomenon and as the Web matures, this correlation could increase.

Finally, with respect to other microscopic measures such as clustering coefficient or edge reciprocity (how many of the edges are reciprocal), the Wikigraph seems to have stabilized in the last years (Figure 12). The average clustering coefficient is roughly 0.23 and the fraction of reciprocal edges is about about 0.13.

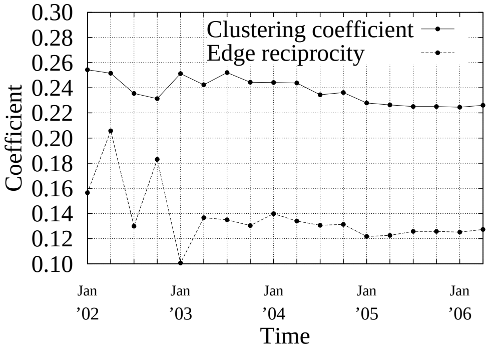

Fig. 12. Clustering coefficient and edge reciprocity over time.

VII. Macroscopic Structure of the Wikigraph

If we disregard the direction of the hyperlinks, we observe that the Wikigraph is almost entirely (weakly) connected. Over 98.5% of the articles are connected to each other.

If we consider the direction of the hyperlinks, the macroscopic connectivity structure of Web graphs can be characterized by mapping their strongly connected components. This analysis generated what is known as the bow-tie structure of the Web [8].

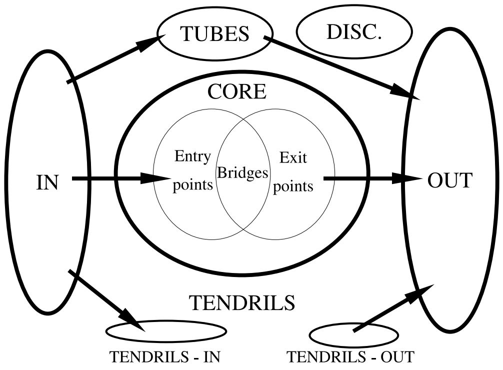

Fig. 13. Bow-tie structure studied by Broder et al. [8]. The arrows indicate link reachability.

This "bow-tie", represented in Figure 13, is formed of four components. The main component is a large strongly connected component named CORE, comprised of all nodes that can reach each other along directed edges. The second and third components are the IN and OUT sets. The IN is the set of nodes that can reach the CORE but cannot be reached from it, whereas the OUT is the set of nodes that are reached by the CORE but cannot reach it. Finally, the set of nodes that cannot reach or be reached from the CORE are the TENDRILS. There are nodes that are reachable from portions of IN or reach portions of OUT. Those TENDRILS that leave a set of nodes from IN and enter a set of nodes in OUT are called TUBES. It can be observed that a significant portion of the nodes are in the large strongly connected component CORE. We can also distinguish sets of nodes completely separated by the main bow-tie, called DISCS.

The first observation about the evolution of this macro structure in the Wikigraph is that the CORE component is getting larger, as can be seen in Figure 14. Currently, about 2/3 of the nodes belong to the larger strongly connected component of the Wikigraph, which is larger than was observed in 1999 in the Altavista crawl [8] or in 2001 in the WebBase crawl [9] but is consistent with measures in more recent samples of the Web such as [24], [25], [26].

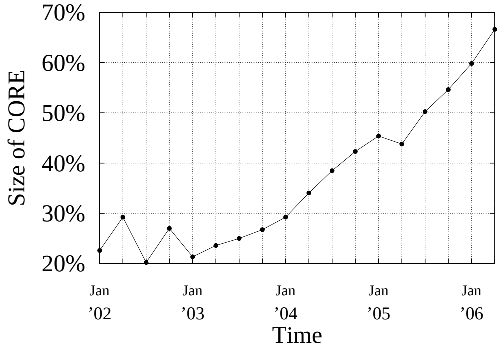

Fig. 14. Relative size of the CORE component with respect to the rest of the graph.

In the past we have observed that in the CORE there may be nodes that are only connected by two links ("unstable" members of the CORE). These nodes can be from 4% to 6% [26], [25] in large Web graphs, but in the case of Wikipedia they are less than 1.5%, meaning that the largest strongly connected component of the Wikipedia is more tightly knitted. We can conclude that the link structure of Wikipedia is well interconnected, in the sense that most of the nodes are in the core, and from any page it is possible to reach almost any other. This is probably due to an implicit aim of an online encyclopedia, that is driving the reader to related topics on the same encyclopedia during the topic description. In this way the content of each article can be fully understood while the surfer visits many different articles.

The increase of the size of the CORE has been mostly at the expense of the OUT component, as shown in Figure 15. The OUT component contains articles reachable from the main strongly connected component but that do not link back to it. We can explain at least in part the reduction of the size component if we consider that the number of out-links per page is becoming larger, as presented in the previous section.

The OUT component of the Wikipedia is also very thin, and over 99% of its nodes are directly reachable from the CORE across all snapshots. The same happens with the IN component, in contrast with larger graphs that exhibit several levels in these components.

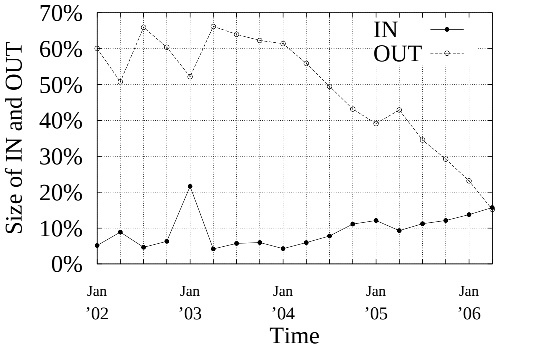

Fig. 15. Relative size of the IN and OUT components with respect to the rest of the graph.

Finally, we study the "movement" of articles among components of the different components of the Wikipedia graph, following [27]. This is depicted by the state diagram in Figure 16 in which only CORE, IN and OUT are depicted, and to each arc we have associated the probability that that a state change occurs from one snapshot to the other.

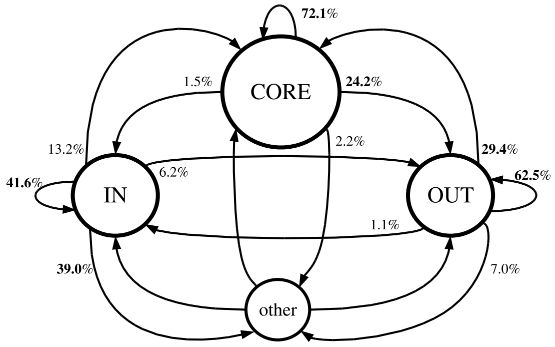

Fig. 16. Migration of articles among the components.

Interestingly, the CORE and OUT components are very stable, with over 60% of the articles remaining in the same component after three months. Also, these transition probabilities have remained basically stable over different snapshots (in the figure we depict the average transition probability across snapshots).

VIII. Concluding Remarks

In this paper we presented a link and temporal analysis of the Wikigraph. We performed a series of measurements and observed that the Wikigraph resembles many characteristics of the Web graph. The core of this study was the temporal analysis of Wikigraphs, where we made a large number of experiments on the evolution over time of the topological and statistical properties of Wikigraphs and made several observations on the frequency of update of the articles of Wikipedia.

The observation of the Wikipedia provides mixed signals of growth and maturity of this collection.

Signs of transient regime (growth):

- The number of articles, updates, visitors and editors is still growing exponentially.
- The size of articles is still growing linearly.
- The number of links per article is also growing linearly, slowly than the amount of text.
- The number of reverts is growing slowly, which may signal more vandalism, but the number of double reverts (revert wars) has stabilized.

Signs of permanent regime (maturity):

- There is a clear power-law distribution of the indegree and outdegree.
- The average edits per user has been mostly constant in the last two years.
- There is a high correlation between PageRank and indegree, indicating that the microscopic connectivity of the encyclopedia resembles its mesoscopic properties.
- The clustering coefficient and edge reciprocity of links have remained basically constant during the last two years.
- Over 2/3 of the articles belong now to the larger strongly connected component.

These are the first observations with this degree of detail of the evolution of a large hyperlinked corpus. In the future, we expect to relate this study with the observed evolution of large samples of pages from the Web.

IX. Acknowledgements

This work was partially supported by the FET Open Project IST-2001-33555 (COSIN), and by the EU IST-FET project 001907 (DELIS).

References

[1] A. Gulli and A. Signorini, "The indexable Web is more than 11.5 billion pages," in *Poster proceedings of the 14th international conference on World Wide Web*. Chiba, Japan: ACM Press, 2005, pp. 902–903. [Online]. Available: http://www.di.unipi.it/%7Egulli/papers/f692_gulli_signorini.pdf

[2] A. Ntoulas, J. Cho, and C. Olston, "What's new on the web?: the evolution of the web from a search engine perspective," in *Proceedings of the 13th conference on World Wide Web*. New York, NY, USA: ACM Press, May 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, May 1999. [Online]. Available: http://www.amazon.co.uk/exec/obidos/ASIN/020139829X/citeulike-21

[4] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer, "Trend detection through temporal link analysis," *Journal of the American Society for Information Science and Technology*, 2001.

[5] *On the bursty evolution of blogspace*. ACM Press, 2003. [Online]. Available: http://dx.doi.org/10.1145/775152.775233

[6] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, October 1999.

[7] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Computer Networks*, vol. 31, no. 11–16, pp. 1481–1493, 1999. [Online]. Available: citeseer.ist.psu.edu/kumar99trawling.html

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web: Experiments and models," in *Proceedings of the Ninth Conference on World Wide Web*. Amsterdam, Netherlands: ACM Press, May 2000, pp. 309–320. [Online]. Available: http://www9.org/w9cdrom/160/160.html

[9] D. Donato, L. Laura, S. Leonardi, and S. Millozzi, "Large scale properties of the webgraph," *European Physical Journal B*, vol. 38, pp. 239–243, March 2004. [Online]. Available: http://dx.doi.org/10.1140/epjb/e2004-00056-6REF:b03602

[10] J. Cho, "The evolution of the web and implications for an incremental crawler," in *Proceedings of 26th International Conference on Very Large Databases (VLDB)*. Cairo, Egypt: Morgan Kaufmann Publishers, September 2000, pp. 527–534.

[11] Z. B. Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, "Sic transit gloria telae: towards an understanding of the web's decay," in *Proceedings of the 13th conference on World Wide Web*. New York, NY, USA: ACM Press, May 2004. [Online]. Available: http://www.www2004.org/proceedings/docs/1p328.pdf

[12] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the Twelfth Conference on World Wide Web*. Budapest, Hungary: ACM Press, 2003.

[13] R. Kraft, E. Hastor, and R. Stata, "Timelinks: exploring the link structure of the evolving Web," 2003. [Online]. Available: http://www.soe.ucsc.edu/∼rekraft/papers/workshop.pdf

[14] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, "Preferential attachment in the growth of social networks: the case of wikipedia," Feb 2006. [Online]. Available: http://arxiv.org/abs/physics/0602026

[15] F. Viegas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of SIGCHI*, Vienna, Austria, 2004, pp. 575–582. [Online]. Available: http://citeseer.ist.psu.edu/700564.html

[16] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, December 2005. [Online]. Available: http://dx.doi.org/10.1038/438900a

[17] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," in *Proceedings of the International Conference on Information Quality*, 2005. [Online]. Available: http://www.isrl.uiuc.edu/∼stvilia/papers/quantWiki.pdf

[18] S. Millozzi, D. Donato, L. Laura, and S. Leonardi, "Cosin tools: a library for generating and measuring massive webgraphs," 2003. [Online]. Available: http://www.dis.uniroma1.it/∼cosin/

[19] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose, "Implicit structure and the dynamics of blogspace," in *Workshop on the Weblogging Ecosystem*, New York, NY, USA, May 2004.

[20] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD '05* New York, NY, USA: ACM Press, 2005, pp. 177–187. [Online]. Available: http://dx.doi.org/10.1145/1081870.1081893

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the Web," Stanford Digital Library Technologies Project, Tech. Rep., 1998. [Online]. Available: http://citeseer.ist.psu.edu/page98pagerank.html

[22] L. Becchetti and C. Castillo, "The distribution of PageRank follows a power-law only for particular values of the damping factor," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2006, pp. 941–942. [Online]. Available: http://dx.doi.org/10.1145/1135777.1135955

[23] G. Pandurangan, P. Raghavan, and E. Upfal, "Using Pagerank to characterize Web structure," in *Proceedings of the 8th Annual International Computing and Combinatorics Conference (COCOON)*, ser. Lecture Notes in Computer Science, vol. 2387. Singapore: Springer, August 2002, pp. 330–390. [Online]. Available: http://www.cs.purdue.edu/homes/gopal/prankfinal.pdf

[24] R. Baeza-Yates, C. Castillo, and E. Efthimiadis, "Characterization of national web domains," *To appear in ACM TOIT*, 2006. [Online]. Available: http://www.chato.cl/research/

[25] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone, "A comparison of sampling techniques for web characterization," in *Workshop on Link Analysis (LinkKDD)*, August 2006.

[26] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas, "Mining the inner structure of the web graph," in *Eigth international workshop on the Web and databases WebDB*, Baltimore, USA, June 2005. [Online]. Available: http://webdb2005.uhasselt.be/papers/P-9.pdf

[27] R. Baeza-Yates and B. Poblete, "Dynamics of the chilean web structure," in *Proceedings of the 3rd International Workshop on Web Dynamics*, New York, USA, 2004. [Online]. Available: http://www.dcs.bbk.ac.uk/webDyn3/webdyn3_proceedings.pdf