Traffic in Social Media I: Paths Through Information Networks

Jacob Ratkiewicz, Alessandro Flammini, Filippo Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing, Indiana University, Bloomington
http://cnets.indiana.edu

***Abstract*—Wikipedia is used every day by people all around the world, to satisfy a variety of information needs. We cross-correlate multiple Wikipedia traffic data sets to infer various behavioral features of its users: their usage patterns (e.g., as a reference or a source); their motivations (e.g., routine tasks such as student homework vs. information needs dictated by news events); their search strategies (how and to what extent accessing an article leads to further related readings inside or outside Wikipedia); and what determines their choice of Wikipedia as an information resource. We primarily study article hit counts to determine how the popularity of articles (and article categories) changes over time, and in response to news events in the English-speaking world. We further leverage logs of actual navigational patterns from a very large sample of Indiana University users over a period of one year, allowing us unprecedented ability to study how users traverse an online encyclopedia. This data allows us to make quantitative claims about how users choose links when navigating Wikipedia. From this same source of data we are further able to extract analogous navigation networks representing other large sites, including Facebook, to compare and contrast the use of these sites with Wikipedia. Finally we present a possible application of traffic analysis to page categorization.**

I. Introduction

The online collaborative encyclopedia Wikipedia has become a mainstream information resource for people worldwide, and is often cited as an example of the power of social media. Wikipedia's more than 7 million articles attracted more than 35 billion hits in 2009, with a few gathering most of the attention (e.g., The Beatles, with 37 million hits) while the majority of pages remain fairly obscure. One may assume a variety of reasons why people uses Wikipedia, e.g., to learn more about news items, satisfy specific curiosities, or carry out research for school assignments. Here we seek to develop an understanding of the motivations and modus operandi behind these varied patterns of access, mostly through the analysis of traffic data. Even a simple inspection of traffic patterns may reveal who is accessing a given Wikipedia article and why. Pages like "biology," for example, show a clear weekly cycle, with lows during the weekends, deeper lows in vacation periods, and peaks during final exam weeks, suggesting that college students are the main users of these pages.

Not infrequently pages show sudden surges in the visits they receive, often in close time proximity to the appearance of related news pieces in other media. In a companion paper [1] we focus on the statistical characterization of 'bursty' events. Here we pay special attention to bursts, in the same way as a physical scientist may pay attention to perturbations of a physical system because they may reveal important facets of its internal dynamics. In particular, we study the correlation of traffic bursts with external events, such as news. Large deviations from 'normal' traffic also offer an excellent probe for how traffic to a page may generate secondary traffic toward linked pages.

Outline and Contributions

This paper examines a number of facets of the dynamics of traffic in, through, and out of Wikipedia, with revealing comparisons to other online systems. This study makes use of a large body of data about human interaction with Wikipedia, which we believe to be unexplored to date. After a discussion on our data sources and the pre-processing and filtering strategies used, we present a suite of analyses which we roughly divide into two types. The first deals with the access to, and egress from Wikipedia; the second delves into the microscopic interactions between users and individual Wikipedia topics. Among the specific contributions of this paper are the following:

- § IV-A discusses an analysis of the general relationship of the Wikipedia network with the Web at large, and explore where Wikipedia users come from, and where they go. Among our findings is the fact that Wikipedia is used both as a encyclopedia, providing links to other pages within itself, and as a directory, providing links to other pages in the Web at large. Motivated in part by this we introduce a graphical technique for visualizing the usage patterns of a network at a high level, able to visually describe the degree to which a network is used like a directory, a search engine, an encyclopedia, or for browsing.
- § IV-B presents a comparison between the traffic patterns observed in Wikipedia and in some other information networks. We find that the usage visualization from § IV-A identifies the intuitive usage patterns of different networks: browsing, encyclopedia, and search.
- § IV-C outlines experiments performed to determine how the bursty nature of popularity for some Wikipedia articles correlates with bursts observed in the Web at large, as measured by Google Trends. Here we find that many bursty Wikipedia pages do correlate with appropriately chosen Trends data, suggesting that these bursts are driven by factors external to Wikipedia, such as the news.

- In § V we take a more detailed look at how traffic moves from page to page inside Wikipedia — the distribution of content similarity of pages which are linked to each other, and what links users preferentially traverse. Some interesting regularities are discovered; traffic between linked pairs of pages is correlated, and often this is the result of traffic flowing from a page to its similar neighbors rather than traffic flowing to neighboring pages from external sources.
- Finally, § VI explores the use of traffic data for predicting categories of Wikipedia pages; here, the goal is to work towards a tool that can suggest to a human author which pages a category might belong to, based on neighborhood, traffic, and content information.

II. Related Work

Web traffic is a proxy for online popularity. While the static properties of Web traffic have been fairly well investigated (e.g., its distribution across all pages or hosts in a given period of time [2]), much less is known about how the traffic toward individual pages changes over time and what factors affect its dynamics, especially when this traffic is characterized by non-regular and intermittent activity. Some prior work on the topic of popularity dynamics has focused on news. Wu and Huberman [3] performed a large-scale study of the news sharing site Digg.com, where users can promote links to articles they like by voting for them. This study tracked the total number of votes that each story receives through its lifetime, finding that this quantity follows a lognormal distribution. They further examined the decay rate of incoming votes for a story, providing insight into the onset and decay of a story's popularity. In general, the dynamics of short-lived events such as the news cycle are relatively well understood; popularity of individual items tends to be distributed according to a lognormal, and stops being accreted after around 36 hours in normal cases [4]. One main difference that sets apart the present study from those mentioned above is that, while attention for a news item is short lived almost by definition, the popularity of a Wikipedia page or topic may be influenced by many news events over an indefinite time span. Therefore the behavior of online popularity cannot in general be characterized by that of individual news-driven events. This is illustrated by considering the difference between the news story "Barack Obama inaugurated as U.S. President," and the Wikipedia article on "Barack Obama." The latter's popularity subsumes that of the former, and of potentially many other news stories.

The popularity of videos on the YouTube video sharing site has been studied by Szabo and Huberman [5] and Crane and Sornette [6]. These dynamics are found to be similar to those of news, but with different popularity classes depending on whether a video has been featured on the front page of the site, or is the type that is likely to be spread by social networks (a so-called *viral* video).

Kleinberg [7] studied the bursts associated with identifiable events in streams, such as the occurrence of a key phrase in a news feed. This approach allows to detect hot topics as temporal bursts in word usage. Kumar *et al.* [8] expanded this notion to analyze the evolution of bursty communities in blogs. On the modeling side Barabasi [9] suggests prioritization as one mechanism leading to bursts of activity. Mathioudakis *et al.* [10] develop a model for attention in social media. Users are viewed as producers of information streams, made of units that may be noticed by other users. This model characterizes items such as blog posts by their 'interaction weights,' a proxy for the degree to which users noticing the items. All of these studies suggest that the dynamics of information access and popularity follows a bursty, intermittent behavior. In a companion paper [1] we propose a simple model to capture some peculiar features of the popularity burst observed in social information networks.

Compared to the existing literature about features of popularity trends such as those mentioned above, work on the potential causes for these trends is scarce. It has been shown that when users have access to popularity rankings (e.g. YouTube views or presence of a book on the New York Times bestseller list), they are more likely to disproportionately favor popular items [4], [6], [11]. Related questions have been explored in the context of the role of search engines in biasing traffic [12], [13].

III. Overview of Data

This section provides a high-level overview of the data sources used in this paper.

A. Wikipedia Article Hits

This data set comes from D. Mituzas, a former software developer for the Wikipedia project who has been logging hits to the Wikipedia proxy server.[1] The data is formatted as compressed text files, one for each hour, containing record tuples of (*language code, article title, hit count*). The data set was initially filtered to retain English Wikipedia pages by considering the language code. 'Special' pages (e.g. talk, image, and user pages) and pages that did not appear in Wikipedia as of June 2008 were filtered out. This latter step was aimed at removing requests that generated a 404 error. Data collection was initiated in February 2008 and continues to the present, although our analysis is restricted to the 13-month timespan between 1st September 2008 and 1st October 2009. For the purpose of this study, this data set has two shortcomings: first it does not contain referrer information, making it impossible to determine where (wiki article or Web page) the visit to a page has originated from; second, it doesn't provides information on what type of agent generated a hit (human or crawler). Figure 1 shows the distribution of the number of hits s received by each page, revealing the same broad features already observed for traffic to Web hosts [2]. In this paper, we refer to this data set as 'page hits,' to distinguish it from our other source of traffic data, to be discussed next.

[1] dammit.lt/wikistats

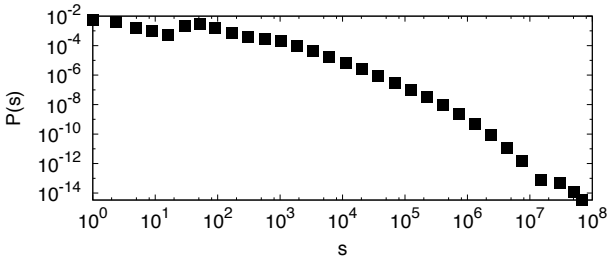

Fig. 1. Distribution of the number of hits received by individual Wikipedia topics, as measured at the Wikipedia proxy server, over the year we observe.

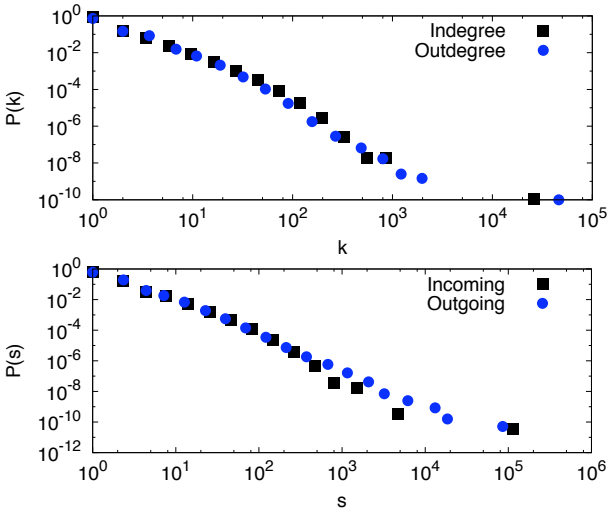

Fig. 2. Distributions of degree (top) and traffic (bottom) for the graph induced by the Indiana University traffic data.

B. Indiana University Web Traffic

A second data set, due to Meiss [2], is a log of Web requests outgoing from all of Indiana University. This data set includes records of the (anonymized) Web browsing activities of about 100,000 faculty, staff, and students of Indiana University from March 2008 to October 2009. The data consists of tuples of the form (*timestamp, agent type, http referrer, target host, target path*). We retained only those (about 5 million) tuples representing requests made by a browser (not a crawler), and involving a Wikipedia article as either a referrer or a target. The chief advantages of this data set with respect to the hits data are that it allows us to exclude automated requests, to track *where* the visitors to pages come from, and where they go when they leave. Its disadvantage is that it represents the actions of a much smaller (and possibly biased) population (Indiana University users versus Web users at large). Figure 2 shows the degree and traffic distributions for the graph induced by this data. We refer to this data set as 'traffic.'

C. Google Trends

Google publishes data about search trends it observes at trends.google.com. Given a query, this site will provide tuples of the form (*date, volume*) representing search trends for that query. The *date* given is at the resolution of weeks, and the *volume* represents the relative volume of queries in that week with respect to an average volume for that query. Rate limits restrict the number of queries that can be addressed to Google Trends. We collected Google Trends data for several hundred Wikipedia topics in order to perform correlation with article hits data, as described later.

TABLE I
TOP REFERRING HOSTS FOR WIKIPEDIA ARTICLES.

referring host	share
en.wikipedia.org	44.81%
google.com	33.99%
empty referrer	9.20%
wikipedia.org	3.56%
search.yahoo.com	1.57%
search.live.com	0.72%
bing.com	0.60%
stumbleupon.com	0.27%
search.msn.com	0.23%
ask.com	0.08%
total	**95.03%**

TABLE II
TOP HOSTS REACHED FROM WIKIPEDIA ARTICLES.

target host	share
en.wikipedia.org	69.66%
indiana.edu	6.18%
boost.org	3.74%
dlib.indiana.edu	1.16%
kinseyinstitute.org	1.10%
omrf.ouhsc.edu	0.56%
banknoteworld.com	0.41%
imdb.com	0.37%
cs.indiana.edu	0.32%
jcmc.indiana.edu	0.23%
total	**83.73%**

D. Wikipedia Text

For our analysis involving page content, we used a full dump of all of the most recent revisions of Wikipedia pages as of June 2009. These dumps are available for download from the Wikimedia project,[2] and contain the full text of all Wikipedia articles, as well as user, talk, redirect, and other types of special pages. Due to resource limitations, past revisions of Wikipedia pages are no longer made available for the English-language version. As a consequence, we chose a version dated approximately in the middle of the time period under scrutiny here as representative of the content and link structure for Wikipedia during the entire period. From this corpus we extracted two data sets: the link graph, and a TF-IDF vector-space representation of the text of each Wikipedia page, after removal of stop-words and markup.

IV. MACROSCOPIC PROPERTIES

A. How Users Come and Go

Our first analysis is aimed at understanding how users reach a Wikipedia page and to what extent their visit fulfills their informational needs, or leads to new resources linked

[2]download.wikimedia.org/

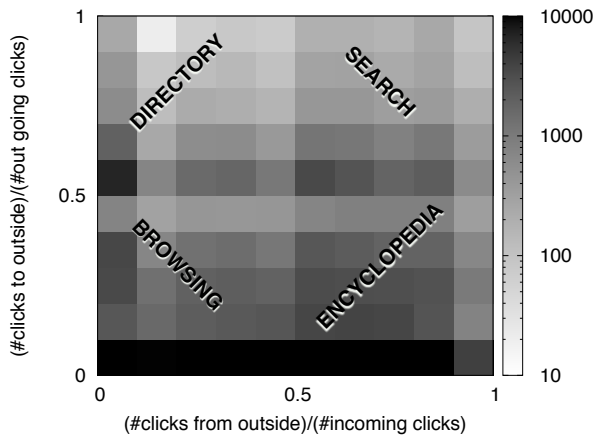

Fig. 3. Usage map of Wikipedia pages. The X and Y axes represent the fraction of a page's traffic that comes from outside, or departs to outside, respectively. The shading of each bin represents, on a log scale, the number of pages with the corresponding usage. See text for an interpretation.

TABLE III
LEAST (TOP) AND MOST (BOTTOM) 'STICKY' CATEGORIES.

title	clicks	stickiness
data structures	5133	$[0.0, 0.02]$
programming constructs	5127	$[0.0, 0.01]$
persistence	5106	$[0.0, 0.01]$
articles with example c code	2991	$[0.0, 0.03]$
stdio.h	2257	$[0.0, 0.02]$
male reproduction	11794	$[0.02, 0.04]$
italian-language operas	2370	$[0.03, 0.06]$
french-language operas	1364	$[0.01, 0.05]$
free software culture and documents	1361	$[0.01, 0.05]$
c headers	1318	$[0.0, 0.04]$
...	...	...
place name disambiguation pages	3149	$[0.97, 1]$
2000s music groups	5584	$[0.93, 0.95]$
grammy award winners	5285	$[0.93, 0.96]$
1990s music groups	4253	$[0.94, 0.96]$
greek mythology	3239	$[0.94, 0.96]$
self-organization	2582	$[0.95, 0.98]$
former british colonies	2241	$[0.92, 0.95]$
1980s music groups	2101	$[0.95, 0.99]$
surnames	1941	$[0.96, 1]$
former spanish colonies	1939	$[0.92, 0.96]$

from the page. When we consider the traffic data for which either the referring or target page is a Wikipedia article, we find that Wikipedia is a traffic sink: the volume of traffic originating from Wikipedia articles (either toward external pages or other articles) is about 30% less than the volume flowing into Wikipedia. Tables I and II show the 10 referring and target hosts for Wikipedia articles that account for the most traffic. The top 10 referring hosts account for 95% of incoming traffic. Our data suggests that most users access articles either from other Wikipedia pages, or are directed there by search engines (mostly Google). About 9% of articles are accessed directly, and the portion of traffic arriving from the rest of the Web is negligible. This documents how Wikipedia has become a well known and relevant resource and is prominently ranked by search engines for a diverse set of queries. About 30% of the traffic originating in Wikipedia is outbound, attesting to Wikipedia's important role as a reference to further information resources. Our data, not surprisingly, show that the traffic to external resources is evenly spread among a large number of hosts, although the specific targets in our data set appear to be strongly biased by our user population. The 70% of internally directed traffic is evidence for the self-referential nature of Wikipedia.

Information about the origin and destination of Wikipedia traffic offers an opportunity to infer the usage mode for specific pages, as shown in Figure 3. This figure displays a heat-map of all Wikipedia pages. Their position along the two axes is determined by the amount of externally originated traffic they receive, and the amount of externally bound traffic they originate. We refer to this kind of plot as a *usage map*. Pages being represented in the upper left quadrant indicate a directory-like usage, with traffic mostly coming from inside and immediately leaving to outside resources. We interpret the upper-right quadrant as 'search' usage (pages visited mainly for the purpose of finding external resources), the lower right quadrant as 'encyclopedia' usage (pages visited from outside and leading to other internal resources), and the lower left quadrant as 'browsing' usage (from one internal page to another). With this interpretation in mind, Figure 3 suggests that while Wikipedia is used in all of these modes, the predominant usage modes are 'browsing' and 'encyclopedia,' as one might expect.

Further insight can be gained by aggregating pages according to their categories. These categories are non-hierarchical topic-descripting tags attributed to pages by their editors. Let us define the average probability that a click originating from a page in a given category will lead to a page inside Wikipedia as the 'stickiness' of the category. We report in Table III the 95% C.I. for the stickiness of the most and least sticky categories. Pages in sticky categories are those responsible for the encyclopedia and browsing usage of Wikipedia. Articles in less sticky categories are mostly used as directory or search pages, to find other resources in the Web at large. We note from this data that Indiana University programmers use Wikipedia mainly as a reference to external pages, unlike people interested in other topics.

B. Comparison with Other Networks

It is informative to compare Wikipedia usage patterns with those of other information networks as done in the previous subsection. To do this, we again leverage our traffic data (§ III-B) by selecting the records whose referring or target host is one of the following:

1) The social networking site Facebook (facebook.com), used by many Indiana University students, staff, and faculty.
2) The Indiana University Knowledge Base (kb.iu.edu), a hyperlinked technical reference site for the IU community that also provides general information of interest to outside users.

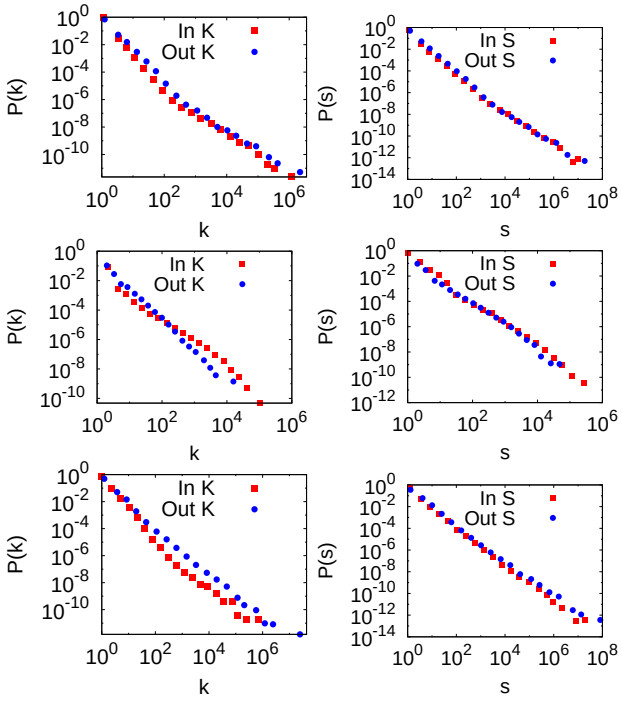

Fig. 4. Distributions of degree (left) and traffic (right) for the Facebook (top), Knowledge Base (middle), and Google query (bottom) networks.

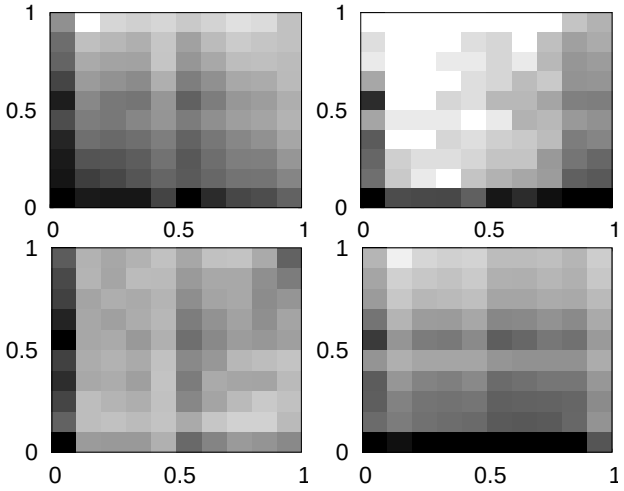

Fig. 5. Usage maps for Facebook (top-left), Knowledge Base (top-right), Google query (lower-left), and Wikipedia (lower-right), visualizing the different modes in which pages in each of these networks are used.

3) The Google search engine (google.com/search).

For each of these sites, we constructed the weighted graph induced by traffic to and from their pages. We ignored requests for subordinate elements like images and advertisements, identified based on file extensions and known ad networks. For the Google and Facebook networks, we removed query strings from all URLs to avoid proliferation of seemingly unique URLs. Figure 4 shows the distributions of node degree and traffic (to or from a node). The largest network is Facebook, followed by Google and the Knowledge Base. In all cases we find very broad distributions of degree and traffic, in agreement with many studies that have reported analogous properties for the Web at large [2].

Figure 5 shows the usage maps for the three networks mentioned above and, for comparison, Wikipedia. Compared to the latter, we see less encyclopedia and more directory usage in Facebook (from users posting external links), as well as a strong browsing component. We also observe more traffic from Facebook to the rest of the Web than in the other direction. The Knowledge Base is used mostly as a proper encyclopedia, with the majority of outgoing traffic being directed to other internal pages. Finally the usage map for Google is the only one with a clear peak in the search quadrant. These observations suggest that usage maps can be a useful visualization tool for how a Web site channels human attention.

C. What Drives Burstiness?

Beyond the above analysis of Wikipedia traffic, our hits data offer a unique chance to take a step back and explore what may trigger users' interests in the first place. In this section we focus in particular on large deviations from *normal* traffic for specific topics. The peculiar distributions of size and frequency for these traffic bursts are characterized and modeled in a companion paper [1]. It is natural to attribute these bursts of activity to real world events, possibly reflected in the news, that trigger the sudden interest of a consistent number of people in a short time span. The analysis described here aims to test this hypothesis by measuring the correlation between the appearance of news on a specific topic and sudden increases in traffic to the Wikipedia page on that topic. We selected first the 200 most bursty articles, where the 'burstiness' of a page is defined as the ratio of its present to previous-day traffic averaged over the time span of the data. We disregarded pages whose present-day traffic was smaller then a threshold (set to 50 hits) to avoid noise fluctuations in traffic. We then constructed queries for each of these pages by removing stopwords and parenthesized words from the page titles; thus "Joe Wilson (U.S. politician)" became "`joe wilson`," and "Army for the Liberation of Rwanda" became "`army liberation rwanda`." These queries were then submitted to Google Trends, and the resulting search volume saved. It should be noted that this normalization process did not, in all cases, produce meaningful query strings; we refrained from correcting these cases by hand to avoid introducing bias. This process resulted in the construction of 200 Google Trends weekly time series. We then computed the Pearson correlation ρ between each Wikipedia topic's traffic and the Google search volume of its associated query.

The results are shown in Figure 6, combined with those of an analogous experiment focusing on the 200 most visited (rather than most bursty) pages. In this figure, the probability density function of the correlation ρ for the bursty pages clearly shows two peaks; one around zero, representing bursty pages with weak or no correlation with search volume data,

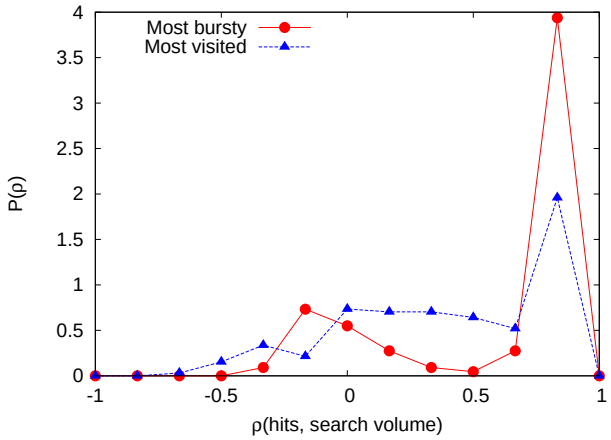

Fig. 6. Probability distribution of the Pearson correlation between the traffic counts of the top 200 most bursty or most visited pages, and their associated search volume from Google Trends data.

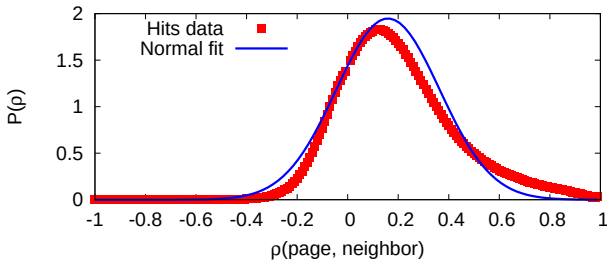

Fig. 7. Distribution of the Pearson correlation ρ between linked pairs of pages at a daily time scale over two months, overlaid with its best normal fit ($R^2 = 0.8$). The discontinuity at 0 represents a high peak.

TABLE IV
MEAN PEARSON CORRELATIONS BETWEEN HITS TIME SERIES.

		All pages	Top 20%
Daily	ρ(page, high neighbor)	0.22	0.52
	ρ(page, neighbor)	0.16	0.29
	ρ(page, random page)	0.04	0.05
Weekly	ρ(page, high neighbor)	0.46	0.68
	ρ(page, neighbor)	0.27	0.35
	ρ(page, random page)	0.25	0.31

and another closer to one representing pages with strong correlation. We hypothesize that the first of these peaks consists of pages that accrete traffic due to internal Wikipedia dynamics; these are explored in the next section. The second peak is clearly due to pages that suddenly receive large amounts of traffic due to news and world events. The distribution of ρ for the most visited pages, however, is more uniform between zero and one. This indicates that popular topics are more weakly correlated with search volume, with a smaller peak around one indicating that people may search for the same sorts of popular things on Google as on Wikipedia.

V. MICROSCOPIC PROPERTIES

We have seen that external events are directly responsible for triggering a large portion of Wikipedia traffic bursts. Let us now explore the dynamics by which users move within Wikipedia, and how they relate to the structure and content of the information network.

A. How Pages Compare

Preliminarily, we examined the Pearson correlation between the time series of hits for pairs of pages satisfying various conditions. Each experiment was duplicated for weekly and daily time resolutions. We found the resulting distributions to be approximately normal; in all discussion below, the normal fits mentioned have $R^2 \geq 0.8$. For example, Figure 7 shows the distribution of the correlation ρ between linked pairs of pages, together with its best normal fit (computed by maximum likelihood). Given these normal distributions, let us compare the traffic correlations by focusing on their means. Table IV reports the estimated means for three hits correlations: between a page and a neighbor (i.e., a page connected by an incoming or outgoing link), between a page and its most correlated neighbor, and between a page and another page randomly selected from the whole Wikipedia. We report the results for the entire data set, as well as for a subset including only the 20% of pages with the most hits. This restricted data set accounts for over 90% of Wikipedia's total hits. All differences are significant at the 99% confidence level, with confidence intervals smaller than the least significant digits shown. We observe that pages are more correlated with their neighbors, and this effect is accentuated when we focus on the top 20% of pages (thus eliminating pages that are visited infrequently). Further, note that the increase in correlation between pages and neighbors versus random pairs of pages all but disappears for weekly time resolution. This indicates that the weekly time scale is so large as to smooth over interesting features in the data; therefore, we omit it in further analysis in favor of the daily time scale.

B. Why Neighbors are Correlated

We now know that neighbors are correlated in the hits that they receive. The observation that two neighboring pages (say a and b) experience similar levels of traffic is consistent with the following two scenarios:

1) Pages a and b are topically similar, and external factors generate interest in their common topic; as a result, both pages experience similar levels of traffic.
2) One of the pages, say page a, sends a large portion of its traffic along its link to page b, causing their levels of traffic to be similar.

To tease apart these effects, we performed several more experiments. The first was to look at at the distribution of content similarity among linked versus random pairs of pages; the results of this experiment are shown in Figure 8. We see that linked pages are far more likely to be similar than randomly chosen ones. When we consider for each page its neighbor with highest hits correlation, we find that the similarity tends to be higher still. Further, we produced a scatter plot representing the relationship between hits correlation and content similarity among linked pages; the result is shown in Figure 9. We find

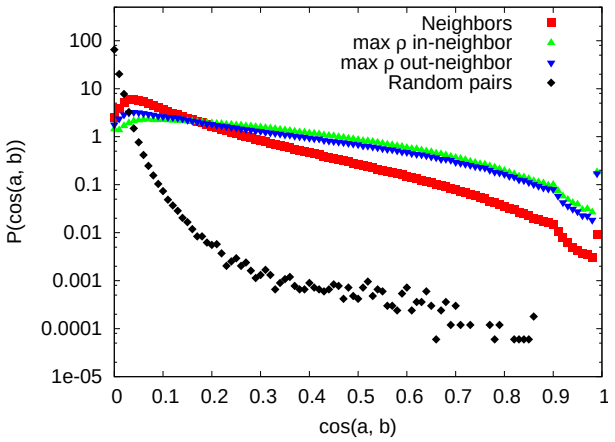

Fig. 8. Distribution of the cosine similarity between pairs of pages selected according to various criteria. The two curves above the neighbor distribution represent the distribution of cosine similarity between a page and specific neighbors; namely, those (in/out) neighbors that have the maximum Pearson correlation (ρ).

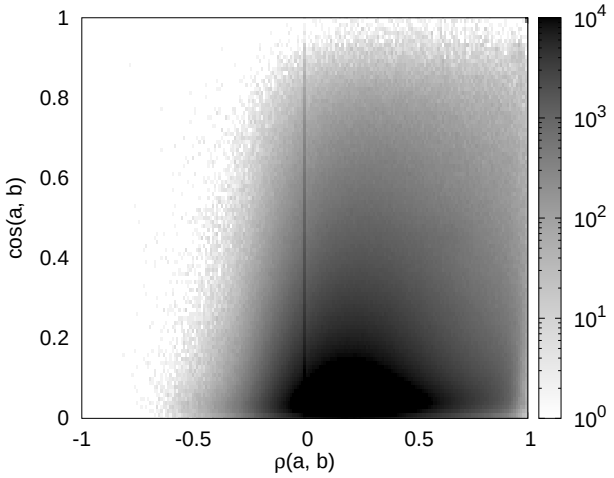

Fig. 9. Heat map visualizing a scatter plot of the the Pearson correlation $\rho(a, b)$ between the hits time series of linked topics a, b at a daily time resolution, versus the cosine similarity $\cos(a, b)$ between their TF-IDF vectors.

that in general, there is a very weak (but non-zero) correlation between the traffic and content similarity of linked pages.

To determine the influence of traffic flowing across links between pages, we need the additional information provided by the traffic data set. We want to see how much of the correlation between the traffic received by two linked pages is due to direct traffic from one to the other. Let $s(a)$ and $s(b)$ be the time series of traffic to topics a and b. Let $s(a \to b)$ be the direct traffic from a to b. Figure 10 shows a scatter plot of the correlation between $s(a)$ and $s(b)$, versus the correlation between $s(a)$ and $s(b) - s(a \to b)$. Points near the diagonal therefore represent pairs of topics whose traffic correlation is not explained by direct traffic between them (scenario 1). Points along the x axis represent pairs of topics whose traffic is no longer correlated when direct traffic is removed (scenario 2). In our data set, this latter scenario is predominant. In other words, traffic from a to b causes in many cases the correlation in traffic between a and b.

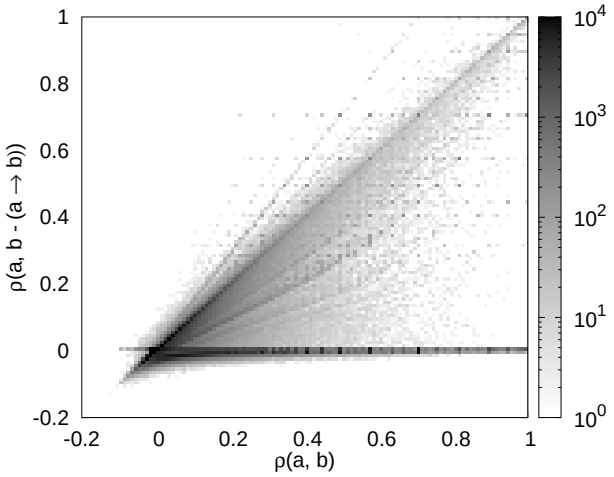

Fig. 10. Heat map visualizing a scatter plot between the Pearson correlation between two page's daily traffic (x axis), and that same traffic when the traffic traveling directly from the first to the second, via a link between them, has been removed (y axis).

VI. Application: Wikipedia Category Prediction

As a potential application of the type of analysis presented here, let us explore some simple techniques for predicting categories of Wikipedia pages — tags assigned by editors. Our task is as follows: for the subset of pages that (a) are in the top 20% of pages by hits, (b) have at least one human-assigned category, and (c) have at least one out-neighbor, use the category assignments of a page's out-neighbors to predict its categories.

Given the category assignment matrix C, where $c_{\chi,p} = 1$ iff category χ has been assigned to page p, we apply a modified nearest neighbors algorithm.

For each page p in our set:

1) Rank p's neighbors by some similarity score (see below). A fraction f of the neighbors will be allowed to vote on p's categories.
2) Let C_p be the union of the sets of categories assigned to each of p's neighbors. Compute a vote weight w_χ for each $\chi \in C_p$ defined as the number of neighbors of p that are assigned category χ.
3) Rank the categories according to the weights w_χ, so that χ_r is the rth category. Evaluate by Mean Average Precision (MAP):

$$\frac{\sum_{r=1}^{|C_p|} P(r)\ c_{\chi_r,p}}{\sum_\chi c_{\chi,p}}$$

where $P(r)$ is the precision at rank r, i.e., the fraction of the top r predicted categories that are correct.

We experiment with three ranking methods for the neighbors q of page p in step (1) of the algorithm: (i) the cosine similarity $\cos(p, q)$, (ii) the hits correlation $\rho(p, q)$, and (iii) the actual

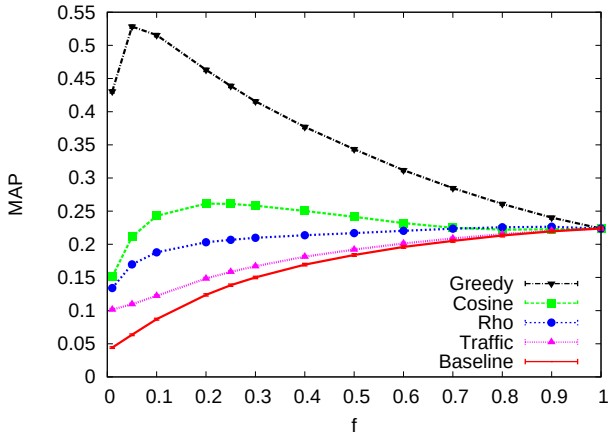

Fig. 11. Mean Average Precision for recovering categories of a Wikipedia page, as a function of the fraction of neighbors allowed to vote for the predicted categories. Error bars for a 95% C.I. are shown, but are so small as to be obscured by the points.

traffic $s(p \rightarrow q)$. Further, to bound the results, we add (iv) a random ranking, and (v) a greedy ranking by the size of the overlap between the category sets of p and q. Note that the algorithm based on ranking (v) assumes knowledge of the categories of p and therefore is not a proper predictor. The results are shown in Figure 11. The method that ranks neighbors by their cosine similarity outperforms all others, achieving a peak MAP for $f \approx 0.2$ before tapering off as less relevant neighbors are added. The methods based on ρ and traffic outperform the baseline, but do not perform as well as content similarity; however, the comparison with the greedy algorithm suggests that all algorithms could be improved. We leave as a topic for future research the question of how to combine these ranking methods to improve their performance.

VII. Summary

This paper presents the results of a major longitudinal study of Web traffic data, across several sites and gathered from several sources. The data are combined to provide a synthesis of Wikipedia usage by real Internet users. Our approach allows for the development of a high-level understanding of the position Wikipedia has with respect to the Web at large; where users come from, and where they go. Further, we introduce a simple graphical visualization (the *usage map*) capable of giving a high-level picture of how pages in a network tend to be used, providing us with a key to interpret the way in which the network itself is navigated. This visualization makes precise some intuitions about how, for instance, the usage of pages on Wikipedia differs from that of pages on Facebook. Further, we find that pages that experience sudden bursts of traffic in Wikipedia often correspond to topics that have attracted sudden bursts of attention in the Web at large, as measured by Google search volume. Results from a number of experiments addressing how users move between pages in Wikipedia are presented. We conclude that users tend to move between pages in some correlation with their content similarity, and that high traffic correlation among neighbor pages is often caused by direct traffic between them. Finally, we tried to exploit similarity in content and traffic among topics to predict Wikipedia page categories. Methods based on traffic fail to outperform those based on content, but there is plenty of room for improvement even in content-based methods; in future work we plan to explore ways of combining these methods.

References

[1] J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Traffic in Social Media II: Modeling Bursty Popularity," under review.

[2] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, "Ranking Web sites with real user traffic," in *Proc. 1st ACM International Conference on Web Search and Data Mining (WSDM)*, 2008, pp. 65–76. [Online]. Available: http://doi.acm.org/10.1145/1341531.1341543

[3] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 45, pp. 17 599–17 601, 2007. [Online]. Available: http://www.pnas.org/content/104/45/17599.abstract

[4] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, and A. Barabasi, "Dynamics of information access on the Web," *Phys. Rev. E*, vol. 73, p. 066132, 2006.

[5] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," arXiv:0811.0405v1 [cs.CY], Tech. Rep., 2008.

[6] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 41, pp. 15 649–15 653, 2008. [Online]. Available: http://www.pnas.org/content/105/41/15649.abstract

[7] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," in *Proc. 12th International World Wide Web Conference*, 2003, pp. 568–576.

[9] A.-L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, pp. 207—211, 2005.

[10] M. Mathioudakis, N. Koudas, and P. Marbach, "Early online identification of attention gathering items in social media," in *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*. New York, NY, USA: ACM, 2010, pp. 301–310.

[11] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, vol. 311, no. 5762, pp. 854–856, 2006.

[12] J. Cho and S. Roy, "Impact of search engines on page popularity," in *Proc. 13th intl. conf. on World Wide Web*, S. I. Feldman, M. Uretsky, M. Najork, and C. E. Wills, Eds. ACM, 2004, pp. 20–29. [Online]. Available: http://doi.acm.org/10.1145/988672.988676

[13] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Topical interests and the mitigation of search engine bias," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 34, pp. 12 684–12 689, 2006.