Change and External Events in Computer-Mediated Citation Networks: English Language Weblogs and the 2004 U.S. Electoral Cycle *

Carter T. Butts[†] B. Remy Cross[‡]

September 11, 2009

Abstract

This study examines global patterns of stability and change within six longitudinal samples of English-language weblogs (or "blogs") during the 2004 U.S. Presidential election campaign. Using distance-based methods of graph comparison, we explore the evolution of the blog-blog citation networks for each sample during the period. In addition to describing the qualitative dynamics of the blog networks, we relate major campaign events (e.g., party political conventions and debates) to the observed pace of change. As we demonstrate, such events are associated with substantial differences in overall network volatility; moreover, volatility is also shown to have strong seasonal and endogenous components. Our findings suggest that external factors (both regular and episodic) may be important drivers of network dynamics.

Keywords: blogs, political networks, dynamic networks, graph comparison, network visualization

1 Introduction

The Internet has provided fertile ground for the study of social networks, attracting interest from researchers across a wide range of disciplines (see, e.g., Carley and Wendt, 1991; Wellman and Haythornthwaite, 2002; Ebel

*This research was supported in part by NSF ITR award #IIS-0331707 and ONR award #N00014-08-1-1015.

[†]Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California, Irvine; Irvine, CA 92697; buttsc@uci.edu

[‡]Department of Sociology, University of California, Irvine

et al., 2002; Monge and Contractor, 2003; Howard and Jones, 2004; Holme et al., 2004). The relative ease of observing and recording interactions in the online environment – as well as the large and increasing volume of computer-mediated communication (Wellman, 2001) – have the potential to provide vast amounts of rich data for network researchers.

Among the many modes of online communication, one that has drawn increasing attention in recent years is the *weblog* (or "blog"), a journal-like hypertext genre which typically combines frequently-updated commentary on current events with hypertext references to primary or secondary information sources. During the months preceding the 2004 U.S. Presidential election, blogs attracted increased interest from political organizations, mainstream media outlets, and academic researchers due to their role as a site for political mobilization (Adamic and Glance, 2005; Howard, 2005; Rainie et al., 2005). As this period represented a critical point in the initial emergence of blogs as a political tool, it is of considerable historical interest. The essentially relational character of blogs, combined with their use in the political process, likewise makes them an obvious target for network research (see, e.g., Backstrom et al., 2006; Hsu et al., 2006; Mislove et al., 2007). The present study is one such effort.

This paper examines global patterns of stability and change within six samples of both politically oriented and non-politically oriented blogs during the 2004 U.S. Presidential election campaign. Using distance-based methods of graph comparison (Butts and Carley, 2001; 2005), we explore the evolution of the blog-blog citation networks for each sample during the period. In addition to describing the qualitative dynamics of the blog networks, we relate major campaign events (e.g., political party conventions and debates) to the observed pace of change. As we demonstrate, such events are associated with substantial variation in overall network volatility. These episodic influences act in concert with seasonal variation in the form of daily and weekly activity cycles, all of which (along with endogenous and sample-specific factors) play a major role in shaping network dynamics.

We begin the paper with a definition of the blog genre, as employed in our study. We then discuss the sampling procedure used to obtain the six groups of blogs to be analyzed, as well as the techniques employed to collect data on these groups over the observation period. This section also offers some comments on issues relating to sampling and data collection for future online studies. After discussing the data itself, we turn to an examination of the pace of change within each sample. Finally, we examine the qualitative dynamics of each network over the observation period.

2 Data

The data for this study consists of six networks of English-language blogs, sampled at six-hour intervals over a period of approximately four months. Before proceeding to our analysis, we describe the criteria used to identify blogs for this study, as well as the sampling methods used to collect the blog data.

2.1 Blog Criteria

The term "blog" is generally used by both participants in and observers of online culture to refer to a family of web sites with a journal-like structure; like most such terms, there is no precise agreement on what does or does not constitute a "blog." Despite the lack of perfect consensus, there seem to be certain basic properties (e.g., chronologically organized text, identified authorship, editorial control by authors) which broadly characterize the genre. For purposes of this study, we employ a fixed set of criteria to define what we mean by "blog"; except as noted otherwise, we use this term exclusively to refer to web sites satisfying this definition.

Formally, our criteria for identification of a web site as a *blog* are as follows:

1. The presence of a clearly marked *front page* identifying the site in question;
2. The presence of multiple *entries* consisting of a heading and text of at least a paragraph in length on the front page;
3. A clear attribution of *authorship* for each entry;
4. A clear *chronology* for each entry;
5. The presence of a specified group of one or more *site authors* responsible for authoring entries on the site; and
6. Evidence of *editorial control* by the site authors over site content.

We note that these criteria exclude certain types of sites from blog status. For instance, sites such as essay archives, news filters, static home pages, electronic versions of most magazines/newspapers, online discussion forums, and the like do not meet our definition of "blog." Such sites do not generally self-identify as "blogs," however, and are not commonly treated as such

by members of the online community. By turns, we are aware of few sites which self-identify and are commonly treated as "blogs" which do not satisfy these criteria; thus, while we recognize that no definition is likely to obtain universal agreement, we posit that our criteria are fairly representative of typical usage within the online community.

To be included in the present study, all web sites were required (based on manual inspection by the authors) to be considered blogs under the above criteria. In addition, our study is limited to blogs which are public and written in the English language.

2.2 Sampling Procedure

There is currently no generally accepted means for sampling from the population of blogs. Most efforts to date have thus been ad hoc, or otherwise limited to a well-bounded population (e.g., all pages hosted by a particular service provider) (Adamic and Glance, 2005; Welsch, 2005). Although the recent development of principled link-trace methods for online social networks (Mislove et al., 2007; Gjoka et al., 2009) holds significant promise for resolving this problem, even these approaches currently founder on poorly connected and/or directed graphs, or in situations for which manual eligibility verification is required. The methods available during the period of data collection for this study (summer/fall of 2004) were even more limited. Because of the complexity of this problem – and the lack of uniform procedures in the field – we thus describe our sampling procedures in detail.

To study the impact of the 2004 US Presidential election on the blog community, it is important to capture relations among several types of sites. First, high-status blogs (e.g., those whose authors were credentialed by the major political parties) are expected to play a major role in election-related communication, and should be included; likewise, it is desirable to have a broader selection of blogs whose content is politically oriented. In order to evaluate change within these groups, however, we also require sites chosen more uniformly from the larger blog population to serve as a point of comparison. The approach taken here is something of a compromise: rather than attempting to obtain a single sample which satisfies all of these criteria, we draw several distinct samples for particular purposes. Although no one sample provides a complete view of the blog population, then, an integrated picture can be developed by comparing results across samples.

The data for this study is derived from six samples. Each of these is described in detail below.

2.2.1 Third-party Designated Samples

The first and most straightforward of the samples considered here are three sets of blogs designated as significant by third parties. The first two samples consist of the sets of blogs whose authors were credentialed by the Democratic National Committee (DNC) and Republican National Committee (RNC) for their respective political conventions. Media credentials provide the recipients with privileged status, as well as enhanced access to restricted events; as such, they are highly sought after at the major conventions, and not readily granted to those outside of mainstream media outlets. The credentialing of blog authors for the 2004 conventions provided one of the first tangible signals of blogs' acceptance as legitimate political entities during the electoral cycle, making this status particularly salient to those obtaining it. Similarly, the selection of blog authors for credentialing in and of itself granted their associated blogs a certain degree of legitimacy as quasi-official outlets for information regarding the election. For these reasons, then, it seems appropriate to single out both the 34 DNC and 14 RNC credentialed blogs as samples of particular interest.

As a counterpoint to those blogs designated as high-status by virtue of political endorsement, it seems also of interest to identify blogs which are particularly prominent within the blog-blog citation network. Such a list was kept by the web service blogstreet.com, which utilized an algorithm similar to PageRank (Page et al., 1999; Brin and Page, 1998) to identify especially "influential" blogs. As PageRank is very similar in function to eigenvector centrality (Koschützki et al., 2005), influential blogs by this measure are those which are deeply embedded within the largest dense cluster of the blog-blog citation network. The blogstreet.com "Top 200" influential blogs (as of 7/18/04) were drawn as a high-status sample for this study; removal of two non-blog sites reduced the sample size to 198. Although (like all such services) blogstreet.com's assessment is based on a limited database of sites, the relative robustness of eigenvector measures to missing data (Costenbader and Valente, 2003) suggests that sites within this sample are likely to be a reasonable proxy for then-prominent blogs more generally. (The tendency of blog tracking services to oversample prominent sites further supports this interpretation.)

2.2.2 Link-trace Samples

One well-known approach to sampling from hidden or hard-to-reach populations is the use of link-tracing designs (Thompson, 1997). In the case of

relational data, link-tracing designs have the added bonus of being ignorable with respect to certain analyses (Thompson and Frank, 2000). Here, we employ link-tracing to draw two parallel samples. The first is a sample of political blogs, where "political" is interpreted to signify having at least 50% of front page content (by area) devoted to electoral, legal, or other expressly political issues. The second is a control sample of arbitrary blogs, drawn in the same fashion as the first. The combination of both samples, then, potentially allows us to control for some effects of the sampling procedure during analysis.

Link-tracing for both samples was performed by a single branch Markov chain in-link design with a maximum depth of 50 and allowing repetition. In particular, each chain was started with a single seed, from which a random in-neighbor was chosen using the service technorati.com. If this site failed to meet criteria it was rejected, and a new in-neighbor drawn at random; once an eligible in-neighbor was found, this site was added to the chain, and the process was repeated. Each chain was terminated when no eligible in-neighbor was available, or when the maximum chain length of 50 sites was reached. A new chain was drawn for each seed site in the sample; once all chains were drawn, the full sample was generated by taking the union of all chains.

The link trace samples for this study were drawn from 7/17/04-7/18/04. 35 seed sites were used in each case. For the control (or "random") link trace, all seeds were uniform draws from the same source as the random blog sample (see below); eligibility criteria were identical to those employed elsewhere in this study. For the political link trace, an ideologically overdispersed set of seed sites[1] was identified by taking the first five eligible blogs arising from each of seven google.com queries (conducted on 7/12/04). These queries were for the terms "conservative," "democratic party," "green party," "liberal," "libertarian party," "reform party," and "republican party" (respectively), combined with the restricting term "(weblog or blog)." Given these seed sites, the above link-tracing method was employed to construct the political trace sample, subject to the constraint that all sites in the trace were required to meet the criteria for inclusion as political blogs. This procedure ultimately resulted in a sample of 392 sites, with 203 sites in the "random trace" sample.

[1] Ensuring dispersal of seed points has well-known benefits for the convergence of multiple Markov chain methods; see, e.g., Gelman and Rubin (1992).

2.2.3 Random Sample

As a final point of comparison, we employ an (approximate) random sample of English-language blogs. Uniform sampling of blogs is a nontrivial problem, due to the lack of a well-validated sampling frame. Here, we attempted to circumvent this problem by querying against a large, continuously updated database of blogs and similar websites, and then removing ineligible sites by rejection sampling. While the resulting sample doubtless deviates to some extent from an exact random sample from the blog population, we expect it to serve as a reasonably good approximation.

The exact procedure employed was as follows. Using the hourly update logs of the web service www.weblogs.com (a site which tracked web pages and RSS feeds from a wide range of sources), we obtained a list of all blogs known to have been updated in the four week period from 6/18/04 through 7/16/04. This yielded an initial sample of 253,815 unique URLs. Uniform draws were then taken from this list (without replacement) and manually checked for eligibility. Approximately 1,050 sites meeting criteria were selected for inclusion of the sample; removal of duplicate/obsolete URLs reduced this number to 1,037. These sites constitute our (approximate) random sample of the blog population.

As noted above, this sample may deviate from a purely uniform sample of blogs to some extent. First, prominent sites, sites with registered RSS feeds, and sites hosted by large blog service providers are more likely to be captured by services such as weblogs.com, and may thus be overrepresented. By turns, obscure, hand-generated sites may be underrepresented. Similarly, sites which had been inactive for more than a month prior to the onset of data collection were not captured by the sampling procedure. While we do not have reason to believe that these biases are severe or consequential for the analyses performed here, it must be borne in mind that all attempts at random sampling of blogs during this period were (and still are) somewhat experimental. Validation and refinement of these techniques continues to be an important problem for research in this area.

2.3 Data Collection

Given the six samples identified above, data collection was performed by automated querying of the specified blog URLs. Each URL was queried at six hour intervals (starting at midnight, Pacific time) for the period from

7/22/04 through 11/19/04, for a total of 484 time points.[2] As there were 1,834 unique blogs among the six samples, 887,656 queries were made in all. Data collected at each site query included all outgoing URLs, as well as certain information regarding front-page content. (No information was collected from internal pages, for multi-page blogs.)

For the purposes of this paper, the data to be analyzed consists of the valued, directed networks of blog-blog citations performed via hypertext reference. The data for each of the six samples may be conceived of as an adjacency array, A, such that A_{ijk} is the number of citations from site j to site k observed at the ith time point. Self-citations are included, and outgoing edges for missing or unreachable sites are treated as missing (see below). Citations included for a given time point are those present on the blog front page at the time in question – thus, edges should be taken as representing the state of the blog as viewed by a site visitor at a given time point, rather than purely novel activity on the part of the initiating site.

Sociograms for the time-averaged citation networks are shown in Figure 1; darker edges indicate higher mean citation rates. As the figure suggests, the six samples vary greatly in terms of both size and structure. While the present paper is concerned with dynamics rather than static structure, the diversity of structural forms obtained through the different sampling methods used here should underscore the importance of attending to design considerations in network studies of blogs or other similar entities.

One obvious advantage of the mode of data collection employed here is that data quality is relatively high; unlike self-report or observational data (Butts, 2003; Bernard et al., 1984), edgewise error is effectively non-existent. Nevertheless, *missing* data can occur due to network or system outages. Such outages may be temporary (e.g., due to a loss of connectivity) or permanent (e.g., due to a site's being withdrawn). As noted above, isolated cases of missing sites are here dealt with by missing data coding of outgoing edges. Overall median dyadic missingness rates by sample are 0 (DNC and RNC), 0.005 (influential), 0.003 (political trace), 0.025 (random trace), and 0.022 (random); in addition, data was unavailable for the 00:00 and 06:00 time points on 10/16 due to a network outage.[3] While attrition appears to account for much of the missing data in the two random samples, its overall

[2] Some additional data was collected prior to and after this interval, but this will not be considered here.

[3] Because credentialing for the RNC convention was not announced until data collection was underway, full data for this sample was not available until 8/13. Since all but four sites were captured by other sampling methods, however, we are able to include data for this network for the entire observation period.

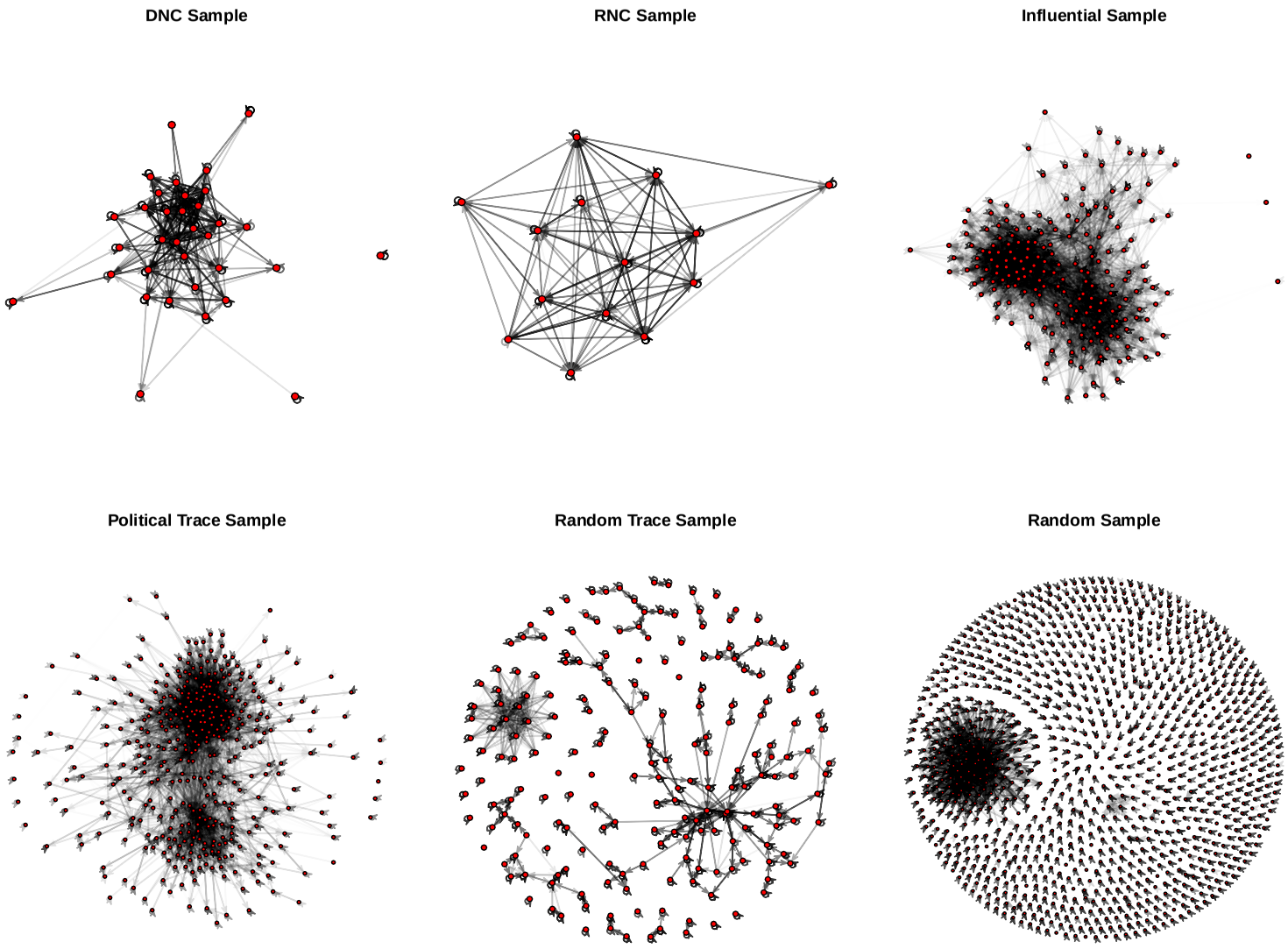

Figure 1: Aggregate Citation Networks, All Samples

extent appears to have been limited: 10 sites were lost from the random trace sample and 38 from the random sample during the four month observation period. The fraction of sites lost to these samples are both below 5%, with other samples losing much smaller fractions. Thus, despite the large number of sites sampled here, we do not encounter high levels of missingness within the data set.

3 Analysis

Given the six blog samples described above, our focus is on the pace and character of change within the blog citation networks during the 2004 US Presidential election cycle. To this end, we employ the distance-based approach of Butts and Carley (2001; 2005), which allows us to conduct an exploratory analysis of change at the network level without having to prespecify particular structural indices to be examined. The picture which emerges from such an analysis is intendedly broad, emphasizing general properties of network dynamics rather than the evolution of specific features. Although we find this approach to yield valuable insights, our use of it here should not be interpreted as obviating further study using low-level analyses.

To capture the differences between networks at various points in time, we employ a normalized L1 metric on the networks' respective adjacency matrices. For adjacency array A on N vertices, the resulting distance is defined as

$$d(i,j) = \frac{1}{N^2} \sum_{g=1}^{N} \sum_{h=1}^{N} |A_{igh} - A_{jgh}| \,. \tag{1}$$

As Butts and Carley (2001) note, the L1 norm is a natural generalization of the Hamming distance for valued graphs, and it shares the twin advantages of weighting all changes equally, and being directly interpretable in terms of edge change units. In this case, the units of d are citation changes per directed dyad. Normalizing changes on a per-directed dyad basis (as opposed to simply counting total changes) facilitates comparison across networks, and also simplifies the treatment of missing data; where directed dyads are unobserved for a given time point, they are removed from both the numerator and the denominator of d.

Utilizing the above distance measure, we construct for each sample a 484 by 484 matrix of network distances, D, such that $D_{ij} = d(i,j)$ (i.e., the distance between the respective networks at time points i and j). By

capturing the relative differences between states at each time point, these distance matrices describe the global evolution of each network over the course of the sampling interval. The D matrices, then, provide the basis for the analyses which follow.

3.1 Volatility and Pace of Change

Before inquiring as to the nature of structural evolution in the blog networks, we begin with the more basic question of volatility. Specifically, it is natural to seek to determine the *pace of change* within each sample, and to examine the extent to which that pace varies during the electoral cycle. Given an intertemporal distance matrix D (as defined above), the net change between consecutive time points (a six-hour interval) is $D_{(i-1)i}$. Given the definition of the elements of D, a natural estimator for the rate of change within the system at times $(2, 3, \ldots, i)$ is $v = \big(D_{12}/6, D_{23}/6, \ldots, D_{(i-1)i}/6\big)$, where the units of v are citations per directed dyad-hour.

Plotting v against time gives us some sense of the extent of volatility among the blog networks. Such plots are shown for each sample in Figure 2. Due to the presence of strong right-skew in the aggregate volatility distribution, the data is shown in log-scale; likewise, a four-day moving average has been superimposed for ease in identifying local trends. As Figure 2 makes clear, the blog networks show no clear global trend in volatility over time. On the other hand, the moving averages show apparent similarities in structure which suggest the possibility of external perturbations.

While external processes could affect volatility among blog networks in a number of ways, one of the most straightforward is by changing the nature and frequency of news events to which blog authors are likely to respond. In the context of the 2004 election cycle, a number of periods suggest themselves as creating distinctive environments for political discussion. These periods (or "epochs") are marked by the beginning and end points of the two major political conventions, the beginning and end of the period in which nationally televised Presidential debates were held, and the beginning and end of Election Day itself. Table 1 lists the nine epochs, along with start times, end times, and numbers of included data points. Under the assumption that the epochs capture the context of political discussion surrounding the national election, we then expect for volatility among political blog networks to be highest during the two conventions, the debate period, and Election Day itself. By contrast, non-political blog networks are not expected to show significant differences by epoch. If such differences are present (and they follow the same pattern as those of the political

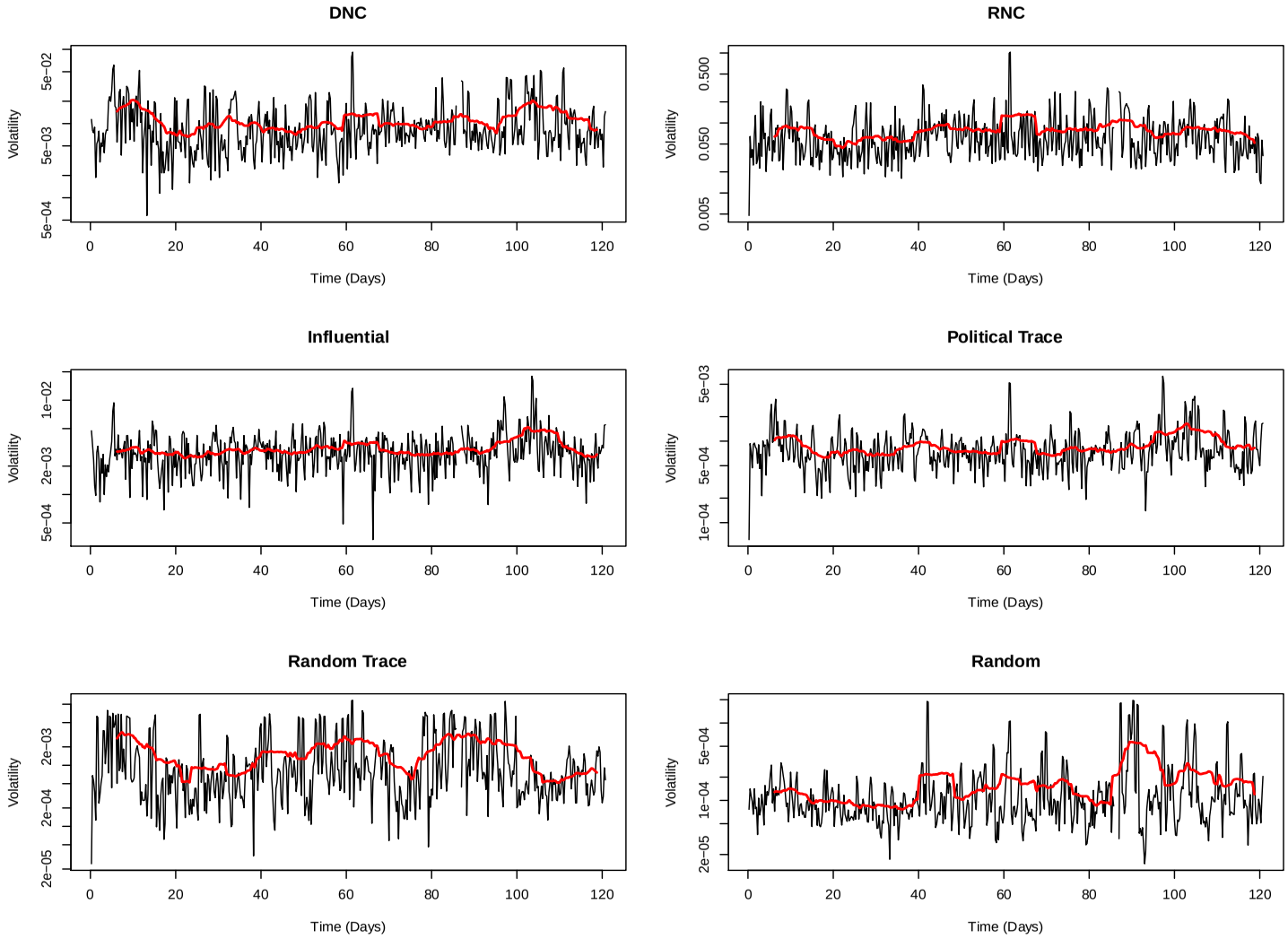

Figure 2: Volatility Over Time, All Samples

blogs), they would suggest a potential influence of the election on interaction beyond explicitly politically interested circles.

In addition to epochal effects, time may exert an exogenous influence on volatility through the rhythms of daily life. Time of day and day of week may both be expected to impact the level of activity within the blog networks, and should be taken account when assessing longer-term trends. Finally, it should be noted that volatility may in fact affect itself, via several potential mechanisms. First, we may imagine a *reactive mechanism,* in which bloggers respond to observed changes on other sites by adjusting their own sites; this would produce positive autocorrelation in v, of an order reflecting the time scale on which bloggers read and respond to one another. Second, we may imagine a *homeostatic mechanism,* in which periods with relatively large numbers of changes tend to be followed by periods of relative quiescence (reflecting an exhaustion of available material), and periods of relative inactivity tend likewise to be followed by periods of relatively high activity (reflecting the adjustments needed to comment on or refer to the "backlog" of relevant topics). Insofar as such a mechanism exists, we would expect it to manifest via the presence of negative short-period autocorrelation in v (net of other factors). Finally, the fact that perturbations of any sort take time to manifest within the blog network may be expected to serve as an *inertial mechanism.* Such a mechanism could operate in two ways: as a positive autocorrelation effect (reflecting inertia in both perturbations and response), or as a positive moving average effect (reflecting only a delay in absorbing perturbations). Taken together, these various mechanisms suggest that blog network volatility may be subject to a combination of period effects (epochal perturbations), seasonality (time/day effects), and serial autocorrelation (endogenous processes).

To capture these effects, we model v within a multivariate ARMA framework (Lütkepohl, 1993). Specifically, we employ a multiplicative model, defining $u_{it} = \log v_{it}$ to be the log volatility for sample i at time t and modeling u via

$$u_{it} = \mu_i + \delta_t + \omega_t + \epsilon_t + \theta_1 u_{it-1} + \theta_2 u_{it-2} + \theta_4 u_{it-4} + \phi_1 z_{it-1} + z_{it}, \quad (2)$$

$$z_{it} \sim N(0, \sigma_i^2) \quad (3)$$

(with z being a white noise process, reflecting idiosyncratic perturbations to the blog system due to news events, individual activities, etc.). μ_i here represents a sample-specific base volatility, δ_t, ω_t, and ϵ_t respectively represent time of day, day of week, and period (epoch) effects, θ and ϕ are autocorrelation (AR) and moving average (MA) parameters, and σ_i is the

Table 1: Epochs in the 2004 Election Cycle

Epoch	Description	Start Time	End Time	Time Points
PreCon	Start of window to DNC Convention	7/22, 00:00	7/25, 18:00	16
DNCCon	DNC Convention	7/26, 00:00	7/29, 18:00	16
InterCon	End of DNC Convention to start of RNC Convention	7/30, 00:00	8/29, 18:00	124
RNCCon	RNC Convention	8/30, 00:00	9/2, 18:00	16
PreDeb	End of RNC Convention to first presidential debate	9/3, 00:00	9/20, 12:00	71
Deb	First presidential debate to last presidential debate	9/20, 18:00	10/14, 18:00	93
PreElec	Post last presidential debate to Election Day	10/14, 00:00	11/1, 18:00	76
Elec	Election Day	11/2, 00:00	11/2, 18:00	4
PostElec	Post election to end of window	11/3, 00:00	11/19, 18:00	67

sample-specific standard deviation of the white noise process. The form of Equation 2 was chosen via a combination of AICC (Hurvich and Tsai, 1989) and residual analysis (specifically, the minimum AICC model for which no residual autocorrelation was obtained), with all estimation performed via maximum likelihood. Sample-specific effects for δ, ω, and ϵ were not found to lead to AICC-preferred models, and these effects were thus treated as homogeneous across samples. Likewise, the selected model exhibits a homogeneous autocorrelation structure, with AR terms at lags 1 (6 hours), 2 (12 hours), and 4 (24 hours), and a single MA term at lag 1. No secular trend was detected. Individual samples did differ substantially in baseline volatility, and in residual variance (σ^2); since most blogs interact with a fairly limited set of peers (see Figure 1), this is largely a density effect. In terms of our putative mechanisms, then, we find support for the impact of both epochal effects and systematic seasonal variation, with additional serial autocorrelation above and beyond these factors. The cross-sample homogeneity in timing effects suggests that – contrary to expectations – the relative impact of these types of exogenous influences is similar across blog samples, although the scale on which those influences operate (driven by μ) varies greatly.

Parameter estimates and goodness-of-fit information for the AICC-selected volatility model is provided in Table 2. (Note: respective temporal reference categories are 12AM, Sunday, and PreCon. Sample intercepts are absolute.) As expected, blog volatility follows a regular diurnal cycle, with activity peaking at mid-day, then falling off with a trough in the early morning hours. Likewise, a clear weekly cycle is present, with a reasonably constant level of work-week activity that declines substantially during the weekend (Saturday being the weekly minimum). The superposition of these two cycles is shown graphically in the top panel of Figure 3, with the dotted reference line reflecting the expected log rate at Sunday, 12AM. Further combining these cycles with the period effects gives a more complete picture of the impact of exogenous timing factors on blog network volatility (Figure 3, bottom). As can be seen, there is a substantial surge of activity during the DNC convention, that largely (but not entirely) subsides until the RNC convention. A modest increase in activity is then observed, which is sustained more or less evenly until the pre-electoral period. At this point we observe a substantial increase in volatility, leading up to a large spike on Election Day. Volatility afterward recedes to levels more characteristic of the RNC convention period, although it does not return to its pre-DNC state before the end of data collection. While the variation in volatility across epochs is often substantial – e.g., the more than two-fold increase from the PreCon epoch to the

Election epoch – the extent of seasonal variation is also striking. Indeed, the pace of change within the blog networks varies almost as much from the 6AM Saturday lull to the 12PM Wednesday peak as it does from PreCon to Election. While the observed level of volatility is a combination of both effects, this serves as an important reminder that network dynamics are no less subject to seasonal variation than other forms of human activity.

Turning to the autocorrelation effects (θ and ϕ), we observe a significant negative AR parameter at lag 1 (6 hours), with positive autocorrelation at 12 and 24 hours. This would appear to be consistent with the presence of both homeostatic and reactive mechanisms, with a 12-24 hour period for bloggers to sample and react to activities in the network (a reasonable lag time, if one assumes that most bloggers during this era were active during 1-2 regular intervals per day). The strongly positive MA effect at 6 hours is greater than the corresponding AR term, suggesting that diffusion of effects due to idiosyncratic perturbations is indeed inertial, while other activity factors are not. Interestingly, the time scale for diffusion due to the MA effect is considerably shorter than the time scale of positive autocorrelation; this is consistent with the notion that the blog networks are fairly quick to absorb new developments, but that the "reverberations" of the resulting discussion (as reflected by endogenous changes in the citation structure) carry on for some time. Such an interpretation would be compatible with the collective problem-solving process identified by Bordia and DiFonzo (2004) in their study of online discussion groups, and the observed pattern of dependence may thus stem from an underlying information processing phenomenon.

Putting this together, then, an analysis of citation volatility over the observation period shows a general main effect of electoral events on the English-language blog population (not merely its politically-oriented subgroups). Periods of major media activity – the two major conventions, Election Day itself (and the immediately preceding period), and the period of the presidential debates – are associated with elevated rates of change relative to the pre-convention baseline. In addition to these period effects, the six blog samples show regular cycles of activity variation on daily and weekly scales, with a magnitude of variability comparable to that of the period effects. Since regular and episodic effects combine to determine expected volatility this underscores the importance of both types of exogenous influences on blog network evolution. Our analysis of volatility also indicates autoregressive and moving average dependence on time scales of 6 to 24 hours, compatible with the combined effects of inertial, homeostatic, and reactive mechanisms. While this analysis cannot tell us exactly how the blog networks evolved, it nevertheless provides clear evidence that this evo-

Table 2: Multivariate ARMA Model, Logged Volatility

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$	
Sample:DNC	-3.3071	0.1155	-28.62	<2e-16	***
Sample:RNC	-1.4067	0.1002	-14.04	<2e-16	***
Sample:PT	-5.6144	0.1477	-38.01	<2e-16	***
Sample:Inf	-4.3722	0.1254	-34.87	<2e-16	***
Sample:RT	-5.5501	0.1592	-34.87	<2e-16	***
Sample:Rand	-7.5095	0.1880	-39.94	<2e-16	***
Time:06AM	-0.2670	0.0252	-10.61	<2e-16	***
Time:12PM	0.2276	0.0335	6.79	1.3e-11	***
Time:06PM	0.1499	0.0267	5.61	2.2e-08	***
Day:Mon	0.2330	0.0483	4.82	1.5e-06	***
Day:Tue	0.1966	0.0517	3.80	0.00015	***
Day:Wed	0.2281	0.0514	4.44	9.4e-06	***
Day:Thur	0.1705	0.0506	3.37	0.00076	***
Day:Fri	0.1337	0.0493	2.71	0.00667	**
Day:Sat	-0.1164	0.0483	-2.41	0.01609	*
Epoch:DNCCon	0.4718	0.1167	4.04	5.4e-05	***
Epoch:InterCon	0.1595	0.0910	1.75	0.07992	+
Epoch:RNCCon	0.2842	0.1194	2.38	0.01736	*
Epoch:PreDeb	0.2584	0.0943	2.74	0.00616	**
Epoch:Deb	0.2430	0.0927	2.62	0.00882	**
Epoch:PreElec	0.4559	0.0935	4.87	1.1e-06	***
Epoch:Elec	0.8506	0.1730	4.92	9.3e-07	***
Epoch:PostElec	0.2775	0.0947	2.93	0.00342	**
θ_1	-0.1071	0.0156	-6.88	7.3e-12	***
θ_2	0.0962	0.0163	5.90	4.1e-09	***
θ_4	0.0490	0.0151	3.24	0.00122	**
ϕ_1^a (Trans)	1.1456	0.0812	14.11	<2e-16	***
σ_{DNC}	0.7159				
σ_{RNC}	0.7041				
σ_{PT}	0.4888				
σ_{Inf}	0.4265				
σ_{RT}	1.1181				
σ_{Rand}	0.7292				

$+\ p < 0.1$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$

[a] Parameter transformed by $(1 - \exp(-\phi))/(1 + \exp(-\phi))$; raw estimate 0.5174

Null deviance: 12545.18 on 2873 degrees of freedom; AICC 12549.18
Residual deviance: 5122.18 on 2841 degrees of freedom; AICC 5188.971
χ^2 (vs. Null): 7423.0 on 33 degrees of freedom; p <2e-16

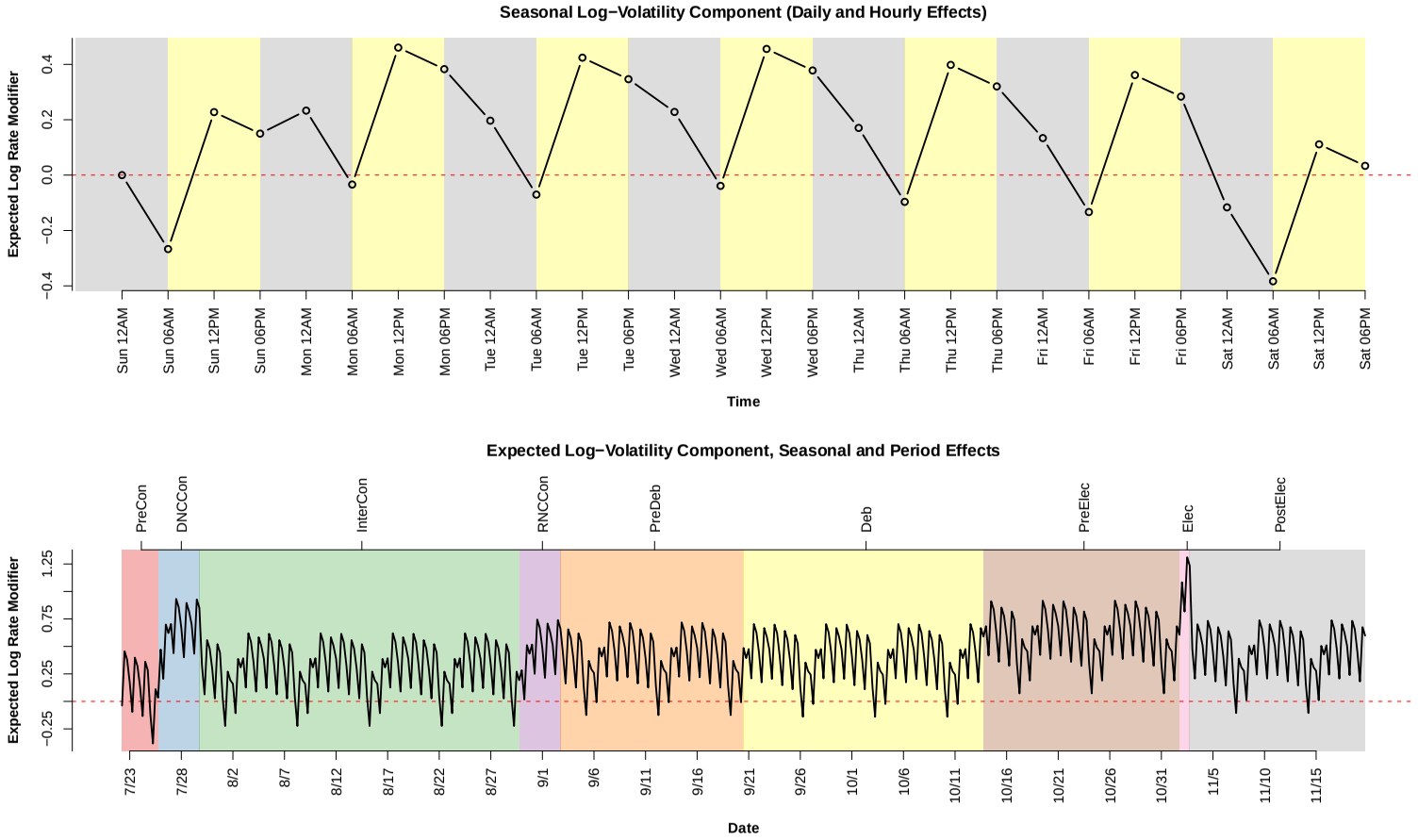

Figure 3: Expected Log-rate Components; Seasonal Effects (Top) and Combined Seasonal, Period Effects (Bottom)

lution was systematically associated with both external and internal social processes.

3.2 Qualitative Dynamics

Although the analyses of Section 3.1 tell us much about the pace of change within the blog samples, they do not inform us as to the *character* of that change. For instance, blog citation networks may evolve as minor perturbations around a baseline structure, as oscillations between several primary configurations, or as cumulative change over time. While the high-dimensional nature of the citation system makes detailed analysis difficult, it is possible to capture the system's qualitative dynamics via multidimensional scaling. Specifically, multidimensional scaling (MDS) methods allow us to represent the differences between system states over time in terms of a set of positions in a low-dimensional, Euclidean space. By examining these latter trajectories, we can then characterize the qualitative behavior of the system as a whole. In the case of the blog citation networks, the inter-state distances are given by D, the matrix of citation differences per directed-dyad for all time point pairs. We begin our discussion of the dynamic content of the D matrix by positing several basic forms which may be reasonably conjectured for the blog citation dynamics, and then proceed to examine the realized forms for each of the observed samples.

3.2.1 Interpreting Trajectories

Given a matrix of inter-state differences across time, a classical metric MDS solution provides us with a sequence of points in a low-dimensional Euclidean space whose respective distances approximate (in a least-squares sense) the original distance matrix. This sequence of points will trace a (possibly irregular) curve, which we refer to as the *trajectory* of the associated system. The system's trajectory, in turn, reveals a great deal regarding the dynamics of the underlying network – in addition to volatility, trajectories allow for the identification of episodic, oscillatory, or other patterns of behavior. To aid in interpreting realized trajectories for the blog networks, we first consider several hypothetical scenarios. In each case, we discuss the dynamics involved, and show the form of the resulting trajectory. As we shall see, only a small number of trajectory features are needed to describe fairly complex system behavior.

First, we may consider the possibility that the blog network tends strongly towards an attractive state, with occasional deviations due to chance events

or external perturbations. Such a state of affairs is illustrated in Panel A of Figure 4; as the Figure indicates, the dynamics of a perturbed equilibrium state should give rise to a "star-like" pattern of random movement around a cluster of nearly identical positions. A second possibility is that the blog network tends to occupy one of a small number of basic configurations at any given time, transitioning back and forth among these configurations in response to internal or external influences. An example of such dynamics is shown in Panel B of Figure 4, for the special case of movement between two base configurations. Note that this transition behavior could be random, periodic, or quasi-periodic; likewise, motion among more than two basic configurations is possible.

Both of the scenarios shown in the Panels A and B of Figure 4 are centered on the idea of essentially transient change: while brief excursions or movement among a fixed set of states are possible, the long-run behavior of the blog network is dominated by a relatively small number of configurations. An alternative possibility is that change is fundamentally *cumulative*. For instance, consider the case in which the blog network gradually evolves through the accumulation of small changes (random or otherwise), such that new changes add to (and do not generally reverse) those which have gone before. This scenario leads to dynamics of the sort shown in Panel C of Figure 4. Here, the long "strand" of states is generated by the accumulated change within the system, in contrast with the "doubling-back" behavior observed in previous scenarios. While the dynamics of Panel C show fairly gradual and consistent change, another alternative is cumulative change which is *episodic*. In this case, the blog network tends to remain in relatively stable configurations for extended periods of time, with occasional periods of rapid change. This scenario is illustrated in Panel D. As the panel shows, episodic change gives rise to an elongated chain of clusters, each corresponding to a period of stability. As with Panel C (but not Panel B), there is no tendency to revisit previous states: change may be irregular, but it nevertheless accumulates over time.

These four scenarios are, of course, ideal cases. One may in principle observe behavior which begins with gradual change but converges to a stable equilibrium, quasiperiodic behavior with local perturbations, etc. Nevertheless, each of these scenarios serves as a basic "building block" whence more complex dynamics may be assembled. By examining the realized trajectory for each blog network, then, we can learn a great deal regarding the qualitative dynamics of the corresponding citation system.

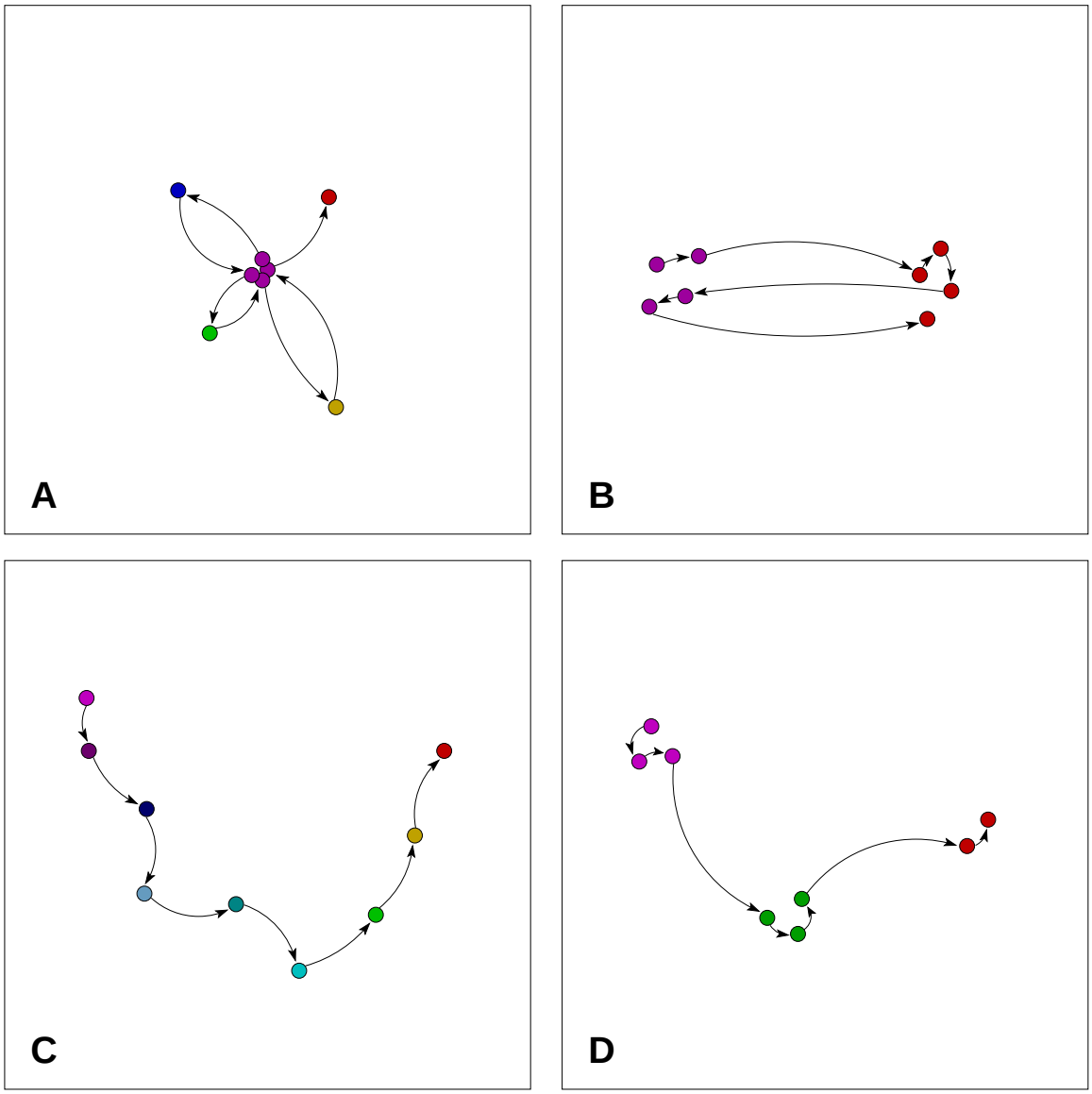

Figure 4: Idealized Trajectories for the blog Citation Networks

3.2.2 Trajectories for the Blog Citation Networks

Having considered the relationship between trajectories and the corresponding network dynamics, we now examine realized trajectories for the six blog citation networks. Two-dimensional metric MDS solutions based on the respective D matrices for each sample are depicted in Figure 5. Individual time points within each panel are displayed using small colored circles; adjacent time points are connected by lines, and the direction of movement is shown by color (with time flowing from blue to red). To connect these trajectories to the epochs of Table 1, major events in the electoral process are also indicated by large colored circles. These reflect, in turn: start/end of the DNC convention (blue); start/end of the RNC convention (red); Presidential/Vice Presidential Debates (grey); and start/end of Election Day (yellow). As per the above discussion, distance between points approximates differences in the corresponding network structures. Thus, proximity indicates structural similarity for the time points in question.

Overall, Figure 5 suggests a fair amount of diversity among trajectories – the patterns observed for the RNC and influential samples, for instance, suggest markedly different underlying dynamics. For purposes of discussion, we will consider the blog networks based on sampling method: the "elite" DNC/RNC/Influential samples; the link trace samples; and the random sample. In each case, we discuss the implications of the observed trajectories for the associated blog dynamics, supplementing Figure 5 with three-dimensional solutions (shown via animations included as linked attachments to this document).

Opinion Leaders: the DNC, RNC, and Influential Blogs We begin with the DNC, RNC, and Influential samples, all of which reflect especially prominent and well-connected sites. Despite this similarity, however, these three networks exhibit quite different trajectories. As Figures 5, 6, and 8 demonstrate, the DNC and Influential samples exhibit a pattern of largely cumulative change over time. This pattern is somewhat smoother for the Influential sample than for the DNC sample, with the latter showing marked changes in velocity associated with the DNC convention and Election Day. Movement for the DNC group seems to stall somewhat in the post-election period and just prior to the RNC convention;[4] interestingly, a sharp turning point occurs in the days following the DNC convention, where the system

[4]Note that *net motion*, as discussed here, should not be confused with volatility per se. It is possible for a trajectory to exhibit high volatility with relatively little net motion, e.g., in the case of random movement within an enclosed region.

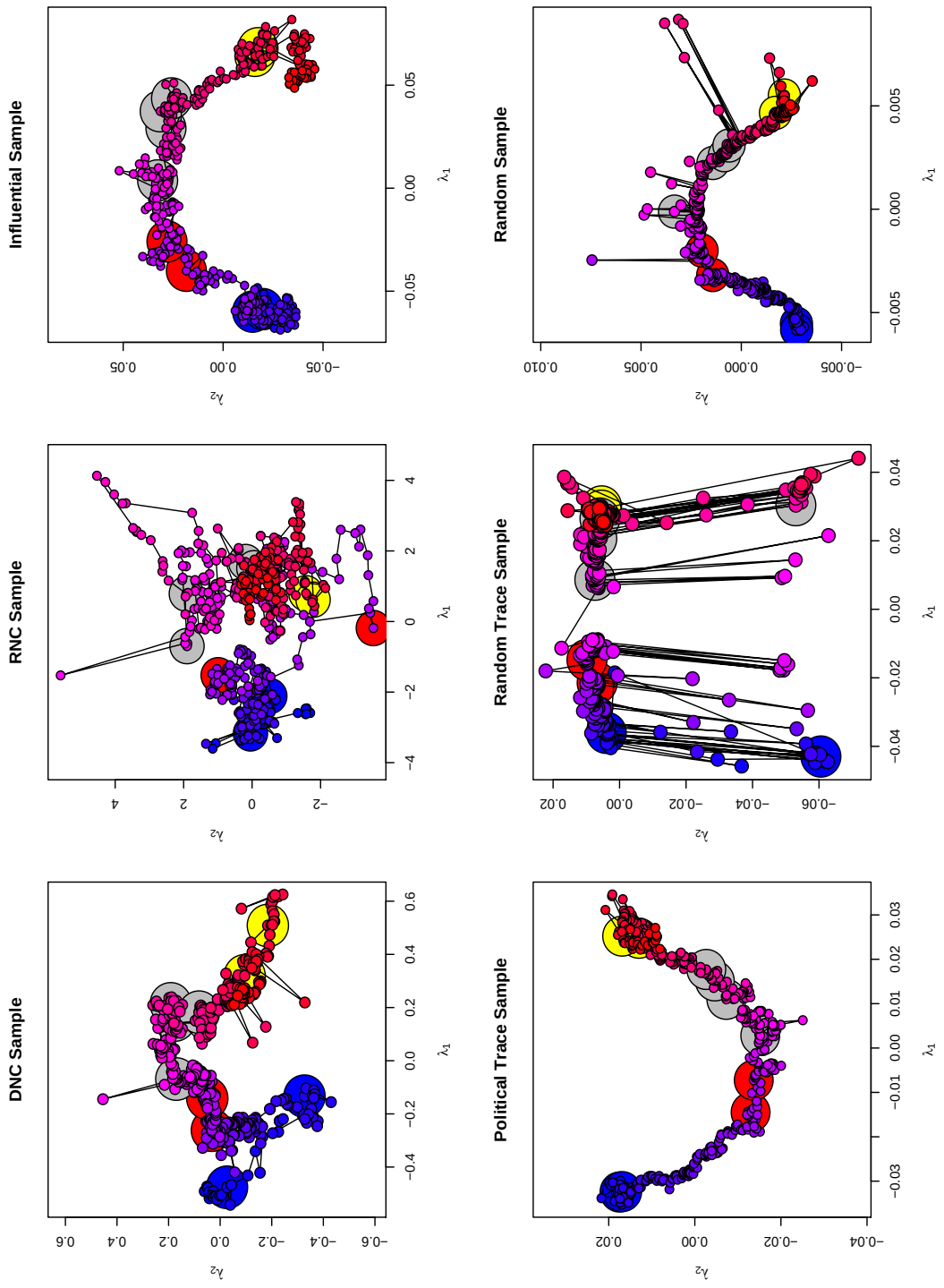

Figure 5: Two-Dimensional Metric MDS Solutions for the blog Citation Network Trajectories

appears to backtrack slightly to a position not far from its pre-convention state. That said, Figure 6 clearly shows that some cumulative change does occur between these two points, and the overall sense of motion is more smooth than episodic.

In contrast with these trajectories, that of the RNC sample (Figures 5 and 7) provides a clear example of episodic motion. The system appears to move erratically within a single region until the period immediately prior to the RNC convention, whereupon substantial net movement is observed. Cumulative change during the convention is then extensive, with high rates of net movement continuing for some days thereafter. At this point, the system seems to settle into a second region, from which it makes occasionally large excursions during the debates, but to which it ultimately returns at the end of the period. In contrast to the DNC sample, relatively little net movement is observed during Election Day itself, although close inspection of the three-dimensional solution shows large perturbations during this period.

<Figure 6: Animated Three-Dimensional Scaling of the DNC Network Trajectory>

<Figure 7: Animated Three-Dimensional Scaling of the RNC Network Trajectory>

<Figure 8: Animated Three-Dimensional Scaling of the Influential Network Trajectory>

Parallel Cases: Political vs. Random Trace Blogs Our next pair of cases consists of the two link-trace samples. While both were collected using the same procedure, they describe two very different populations of blogs (as suggested by Figure 1); perhaps unsurprisingly, they also exhibit very different dynamics. The Political Trace sample (as shown in Figures 5 and 9, shows a fairly consistent pattern of cumulative change over time. While the pace of change does vary somewhat, and a few large perturbations are present (e.g., shortly after the first debate), the overall dynamics are quite smooth. This is a stark counterpoint to the Random Trace sample (Figures 5 and 10), which exhibits an unusual combination of cumulative change and large, temporary perturbations. This trajectory seems to result from the presence of short flurries of intermittent citation within a sample which otherwise exhibits a fairly low level of activity. While there is gradual change in the underlying mean network, this change is slow compared to the frequent "spikes" of activity involving a small number of sites. As was

noted earlier, such profound differences among samples drawn using similar methods underscore the importance of attending to design when attempting to generalize from studies of online networks.

<Figure 9: Animated Three-Dimensional Scaling of the Political Trace Network Trajectory>

<Figure 10: Animated Three-Dimensional Scaling of the Random Trace Network Trajectory>

Baseline Movement: the Random Blogs Finally, we consider the (approximate) Random sample. This network is by far the largest and sparsest of those considered, and its dynamics (Figures 5 and 11) are strongly reminiscent of the random trace sample. While the network shows a clear accumulation of change across time, large local perturbations are also present (reflecting temporary activity in a generally quiescent network). The lower frequency and relative magnitude of these perturbations may be due both to the larger size of the Random sample (which has a smoothing effect) and to the fact that the Random Trace sample was deliberately constructed to oversample edges (and hence activity).

<Figure 11: Animated Three-Dimensional Scaling of the Random Network Trajectory>

4 Discussion and Conclusion

In this paper, we have examined the behavior of six samples of blogs over the months leading up to the 2004 US Presidential election (and its immediate aftermath). Using the metric distance methods of Butts and Carley (2001), we have explored both the pace of change in response to major events, and qualitative aspects of network change over the observation period. Our major findings can be summarized as follows.

Volatility in blog network structure varies substantially over the course of the observation period, and is associated with both political events and seasonal factors. For these samples, mean volatility is positively related to major extended events such as the party conventions and Election Day itself, with slower change occurring during the interstitial periods. Seasonal variation consists of the superposition of a diurnal cycle (with activity peaking at mid-day) and a weekly cycle (with activity falling off on weekends). Thus, both regular and episodic exogenous factors appear to impact blog

network dynamics. Volatility also shows evidence of endogenous behavior, with negative autocorrelation at short time scales giving way to positive autocorrelation on time scales of 12 to 24 hours. This is consistent with the combination of a local homeostatic mechanism (wherein blog activity is related to a constant influx of stimuli that can "accumulate" or be "used up" over short periods) and a longer-range reactive mechanism (wherein bloggers absorb and react to changes from the previous day). A large and positive moving average effect for short time scales also suggests an inertial mechanism, in the sense that new stimuli (e.g., news events or other perturbations) take several hours to diffuse through the system.

In terms of the character of the network dynamics, change within the blog networks appears to consist of either (smooth) cumulative change, cumulative change with strong local perturbations, or episodic change. Smoother change was observed for the DNC, Influential, and Political Trace samples, with large perturbations characterizing the two Random baselines, and the RNC sample showing strongly episodic behavior. Neither periodic/oscillatory behavior nor local variation around a single attractor characterized any of the six samples. For the DNC and RNC samples, net movement was strongly associated with their respective conventions. Although some "backtracking" was seen to occur in the days following the conventions, these events were associated with lasting structural change in both instances. Large local perturbations for the baseline samples appear to result from the relatively low levels of background movement in each case; such behavior is likely to be observed in other large, sparse networks.

While shedding light on some aspects of blog behavior during the 2004 electoral cycle, our study also raises a number of questions. For instance, where cumulative change occurred, it would be useful to know whether this change reflected alterations to the underlying configural properties of the network (its "unlabeled structure"), or whether such change was due primarily to exchanges of positions by particular blogs. Such a question can be examined through the use of structural distances (Butts and Carley, 2005), but the computational difficulties in computing such distances for large graphs are considerable. A related question lies in the observation that aggregate structural change was associated with events such as the major party conventions and the election itself, but not the televised debates. Does cumulative change in a group's internal citation pattern require a temporally extensive event, or is this simply reflective of the relative attention provided to these stimuli? Such a question has important implications for understanding the long-run impact of singular events on online communities.

Beyond these case-specific findings, this study arguably holds a more

general lesson regarding the importance of considering of external perturbations – both episodic and seasonal – when modeling network dynamics. While networks are often said to be "open systems," it is tempting to model them without considering external factors. Depending on one's scientific intent, this may be acceptable; on the other hand, networks such as those studied here clearly *do* respond to external events, and there is a limit to how far we can go in explaining the dynamics of such networks without taking this into account. It is hoped that this study will lead to a more "open" view of network dynamics, both online and elsewhere.

5 References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *LinkKDD 2005: Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining*, pages 44–54. ACM Press, New York.

Bernard, H. R., Killworth, P., Kronenfeld, D., and Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13:495–517.

Bordia, P. and DiFonzo, N. (2004). Problem solving in social interactions on the internet: Rumor as social cognition. *Social Psychology Quarterly*, 67(1):33–49.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: a bayesian approach. *Social Networks*, 25(2):103–140.

Butts, C. T. and Carley, K. M. (2001). Multivariate methods for interstructural analysis. Technical report, Center for the Computational Analysis of Social and Organization Systems, Carnegie Mellon University.

Butts, C. T. and Carley, K. M. (2005). Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory*, 11(4):291–305.

Carley, K. M. and Wendt, K. (1991). Electronic mail and scientific communication: A study of the soar extended research group. *Knowledge: Creation, Diffusion, Utilization*, 12(4):406–40.

Costenbader, E. and Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks*, 25:283–307.

Ebel, H., Mielsch, L. I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E*, 66(035103).

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2009). A walk in Facebook: Uniform sampling of users in online social networks. arXiv.org:0906.0060.

Holme, P., Edling, C. R., and Liljeros, F. (2004). Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174.

Howard, P. N. (2005). Deep democracy, thin citizenship: The impact of digital media in political campaign strategy. *Annals of the American Academy of Political and Social Science*, 597.

Howard, P. N. and Jones, S., editors (2004). *Society Online: The Internet In Context*. Sage, Thousand Oaks, CA.

Hsu, W. H., Weninger, T., Pydimarri, T., and Paradesi, M. S. R. (2006). Collaborative and structural recommendation of friends using weblog-based social network analysis. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*, pages 55–60. AAAI Press.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., and Zlotowski, O. (2005). Centrality indices. In Brandes, U. and Erlebach, T., editors, *Network Analysis: Methodological Foundations*, chapter 3, pages 16–61. Springer-Verlag, Berlin.

Lütkepohl, H. (1993). *Introduction to Multivariate Time Series Analysis*. Springer-Verlag, Berlin.

Mislove, A., Marcon, M., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM Press, San Diego, CA.

Monge, P. R. and Contractor, N. (2003). *Theories of Communication Networks*. Oxford University Press, New York.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Unpublished manuscript.

Rainie, L., Cornfield, M., and Horrigan, J. (2005). The Internet and Campaign 2004. Technical report, The Pew Research Center for the People and the Press: Internet and American Life Project, Washington, D.C.

Thompson, S. K. (1997). Adaptive sampling in behavioral surveys. In Harrison, L. and Hughes, A., editors, *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, pages 296–319. National Institute of Drug Abuse, Rockville, MD.

Thompson, S. K. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26(1):87–98.

Wellman, B. (2001). Computer networks as social networks. *Science*, 293(5537):2031–2034.

Wellman, B. and Haythornthwaite, C. A. (2002). *The Internet in Everyday Life*. Blackwell, Oxford.

Welsch, P. (2005). Drawing water: An ecology of the political blog. Technical report, Association of Internet Researchers, Chicago, IL.