

Abstract

Can internet search activity provide early signals of impending unforeseen events? Things like assassinations, Nobel Prize selections, and terrorist attacks are unexpected, but they often require advance planning; Assassins hatch a plot and the Nobel Committee debates and researches the various candidates. We find that such highly secured secrets often become publicly observable through abnormal search activity on public-access pages. Despite their highly confidential nature, for example, a substantial portion of assassinations show unusually high pre-event search activity. Similar effects are shown for Nobel Prize announcements and secretive military activities. While it is impossible to perfectly predict the future, by aggregating information across individuals, internet search activity may be useful in preventing terrorist attacks, lessening health epidemics, and reducing financial panics.

Using the Internet to Spot Secrets

Yaniv Dover¹, Jonah Berger², Jacob Goldenberg¹, Daniel Shapira³

¹ *Hebrew University, Jerusalem, Israel 91905*

² *University of Pennsylvania, Philadelphia, PA 19104*

³ *Ben-Gurion University, Beer Sheva, Israel, 84105*

Is it possible to anticipate the unforeseen? By aggregating information across individuals, we suggest that internet search traffic can be used in a similar way to prediction markets (1), shedding light on what will happen in the future (2). In particular, search traffic may be especially useful in one place prediction markets fall short. While ethical concerns have arisen about using betting markets to predict assassinations and terrorism (3), practical issues also arise when using markets to forecast secretive events where only a small set of insiders have relevant information. In contrast, we suggest that even highly secured secrets often become publicly observable through search traffic.

Consider two examples. On March 1st, 2008, the Colombian army assassinated FARC Commander Raul Reyes in a highly covert operation, and on October 9th, 2008, Jean-Marie Gustave Le Clézio was announced as the Nobel Prize winner for literature. While both events completely surprised the general public, search traffic for both individuals provide telltale evidence of an impending event. No Spanish Wikipedia page existed for Reyes until the day of his death, but in the week prior, 106 attempts were made to access the non-existent page (Figure 1a). Similarly, even though he was considered a longshot, Le Clézio's Wikipedia page shows intensified activity 10–15 days prior to the prize announcement (Figure 1b). Such unusual activity was not apparent for the other, more likely, candidates (4).

These events were planned by a small set of individuals who undoubtedly wanted to maintain secrecy. They, or more peripheral individuals, also likely want to gather information about the targets, either by talking to others, or searching the Internet. Such attempts left their residue – greater search activity on relevant webpages. Our analysis capitalizes on this activity.

To empirically examine these occasions, we selected a group of events which are planned by a handful of people and where information is highly secured (i.e., political assassinations). We took all assassinations (attempted or successful) from 2008 listed on Wikipedia (4) and used Event Study Analysis (5) to assess whether there was abnormal search traffic before each event. Significant abnormal activity was observed in 36% of the assassinations, and overall there was significant abnormal activity across this set of events ($t(28)=2.82, p<.005$, Fig 1c). It is also worth noting that while some individuals did not receive any pre-event search activity, of those that did have traffic, 71% had abnormal activity before the event.

We further tested the significance of this traffic by comparing it to control events (i.e., accidental deaths). These instances should be similarly unexpected, but should not display abnormal pre-event activity because they do not have planners. Indeed, zero control events demonstrated abnormal pre-event activity ($t(28)=-0.08, p<0.93$ across the set) and overall abnormal pre-event activity was significantly larger in the target set ($t(28)=-2.79, p<.01$).

Though we focused on a bounded set of events for rigorous analysis, similar incidences exist in many other domains. Terms associated with the location of Israel's 2008 destruction of a Syrian nuclear plant also show a dramatic boost in searches right

before the attack and we have found abnormal activity preceding discrete Russian missile tests.

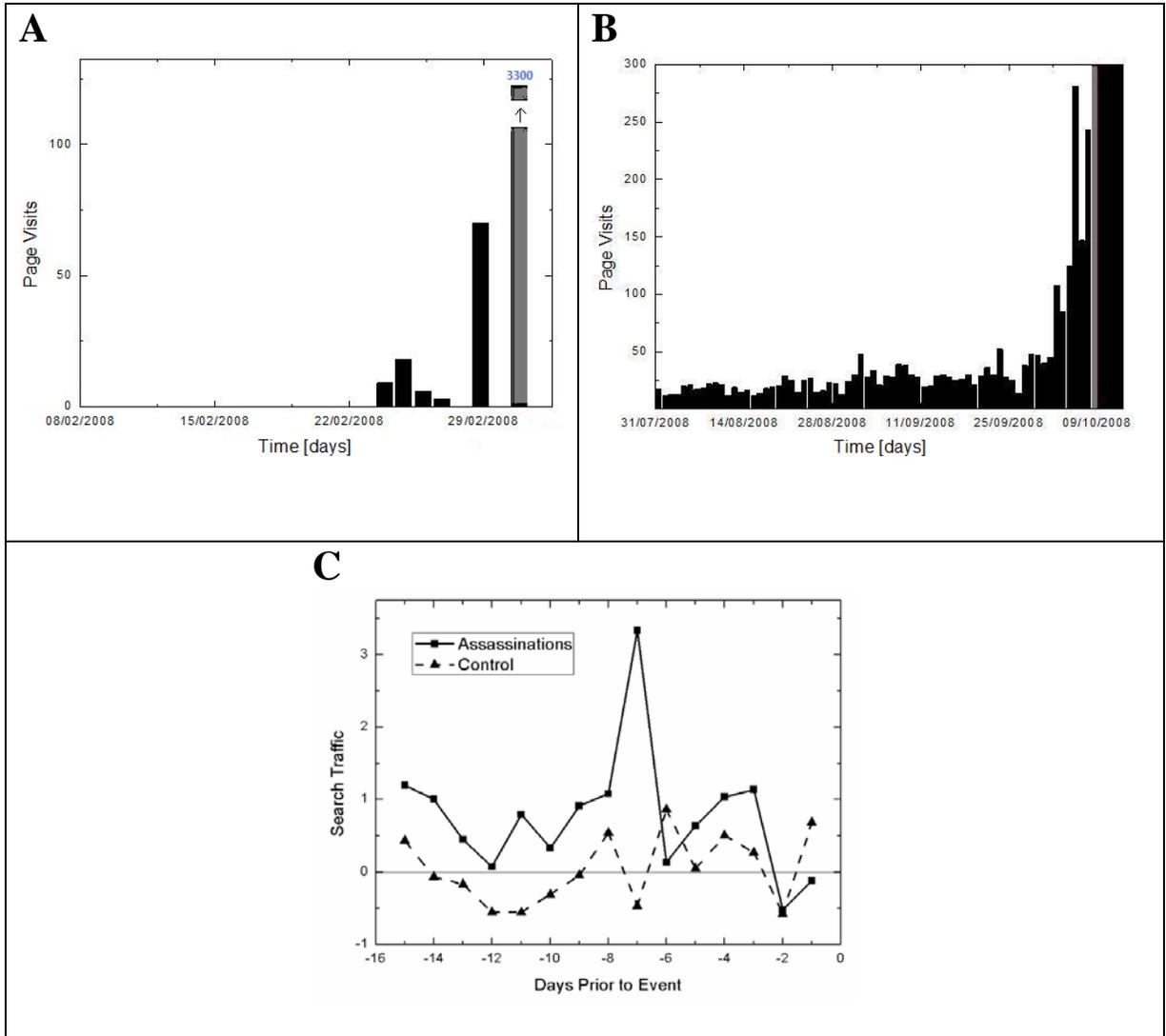
It is interesting to consider who is driving this activity. While central planners may look to the internet for extra background, the activity may reflect the spread of information to peripheral individuals who have little to do with event planning (6). Such outsiders are more likely to lack relevant information, and given the secretive nature of the proceedings, may be driven to covertly learn about potential targets. Regardless of who is searching, however, their actions can act as early signals of the impending event.

These findings have important implications for forecasting unforeseen events. While abnormal search activity does not indicate an exact cause, it suggests unusual interest in a given subject, place, or individual. Applications may range from counterterrorism and public health (2) to mergers and financial forecasting. In conclusion, people privy to secrets have probably always behaved similarly. As the Internet comes to reflect our daily lives, however, it allows us to record, observe, and trace effects that may have been less apparent in the past.

References

1. K. Arrow *et al.*, *Science*. **320**. 877-878, 2008.
2. J. Ginsberg *et al.*, *Nature*. **Nov 19**, 1-3, 2008.
3. R. Looney, *Strategic Insights*, **2**, 2003.
4. Supporting materials
5. J.Y. Campbell, A.W. Lo, A. Craig, *Econometrics of Financial Markets* (Princeton Univ. Press, Princeton, NJ, 1997).
6. L. Sussman, *Bus. Horiz.* **51**, 331-339 (2008).

Fig 1. Daily Wikipedia page visits directly preceding the event (denoted by a gray bar) for a) Raul Reyes, b) Marie Gustave Le Clezio. Panel c) shows aggregate search traffic (across assassinations and accidental deaths, z-scored) for two weeks prior to each event in the main analyses. The bump at 7 days prior to the event is driven by a single event, but our effects remain the same if that event is removed from the dataset.



Supporting Online Materials

We used the Wikipedia page visits statistics tool (<http://stats.grok.se/>) to measure search activity. We took all attempted or successful assassinations from 2008 listed on Wikipedia (S1, S2). We were unable to examine assassinations from earlier than 2008 because search data was unavailable for non-English pages prior to February 2008. It was also unavailable for all pages for the second half of July 2008, so six assassinations were dropped due to lack of data. For control cases, we used Wikipedia lists to collect the same number of accidental deaths (e.g., falling, animal attacks, climbing accidents, traffic accidents, S3). We selected the first cases on each accident list, ignoring any with correlations to other external events or cases where foul play was suspected (to avoid events which might have been planned). Common names with multiple Wikipedia pages for different people could not be analyzed and were dropped.

Google news, Google blogs, and other internet news sites were used to test whether external events or news coverage could be driving abnormal pre-event search activity. No cases showed any significant evidence of external events or pre-event news or blog activity, casting doubt on the possibility that this alternative could be driving our results. In the case of Clezio, for example, while Italian writer Claudio Magris, Syrian poet Adonis and Israeli author Amos Oz, were all leading candidates in the media, Clezio was not mentioned as a possible winner by any major media outlets.

We used Event Study Analysis to test whether activity in an event window is abnormal relative to some baseline “normal” estimation period. We used the simple assumption that:

$$\hat{\varepsilon}_i^* \sim N(0, V_i)$$

Using the notations of (5), where $\hat{\varepsilon}_i^*$ is the abnormal returns vector and V_i is the conditional covariance matrix. We assumed the simplest model, namely the constant mean and variance model for the site traffic and we chose the estimation windows length to accommodate this assumption. For standardization, we used a constant event window (i.e., where pre-event activity might exist) of 15 days preceding the event and the estimation period to be the 35 days before the event window. Similar results were found using window of different lengths. We used t-tests to examine the mean difference of the activity (between the estimation period and event window), and F-test to examine the variance difference between these two windows/time series (both Table S1). Variance tests allow detection of more localized activity (e.g., two of 15 days might show very unusual activity, but would go undetected if only the mean activity across the set of days is used). Cases which exhibited significant abnormality on either measure were considered to show abnormal activity.

References

1. http://en.wikipedia.org/wiki/List_of_people_who_survived_assassination_attempts
2. http://en.wikipedia.org/wiki/List_of_assassinated_people
3. http://en.wikipedia.org/wiki/Lists_of_people_by_cause_of_death

Table 1: Event analysis results for assassinations and control cases. In each group, cases are presented in descending order of significance. For each group of cases the 2nd column contains the mean difference test (t-statistic followed by the mean difference). The 3rd column contains the variance difference test (F-statistic). p-values are noted in parentheses and ** denotes significance by either test at the .05 level.

Assassinations			Control Group (Natural Deaths)		
	Mean Difference	Variance Difference		Mean Difference	Variance Difference
Xanana Gusmão**	2.35, 58.36 (0.022)	41.8 (5.50E-17)	Marion Dewar	2.1, 1.42 (0.16)	1.73 (0.21)
Bill Gwatney**	2.29, 0.175 (0.026)	6.85 (2.00E-06)	Crispin Beltran	-1.18, -0.8 (0.42)	0.45, (0.12)
Jeyaraj Fernandopulle**	1.19, 1.36 (0.23)	2.97 (0.009)	John McWethy	-0.158, -0.23 (0.81)	1.33 (0.46)
Borislav Georgiev**	0.96, 0.11 (0.34)	3.89 (0.001)	Marie-Françoise Audollent	0.025, 0.027 (0.97)	0.49 (0.16)
José Ramos- Horta**	0.8, 1.68 (0.09)	2.46 (0.03)	Abraham K. Biggs	No Internet Activity	
Yaakov Alperon	0.76, 0.09 (0.45)	1.43 (0.38)	Carlos Alinho	No Internet Activity	
Maumoon Abdul Gayoom	0.24, 1.2 (0.81)	0.54 (0.26)	Rolf Bae	No Internet Activity	
Gabriel Mkhumane	No Internet Activity		Ola Brunkert	No Internet Activity	
Mario Fernando Hernández	No Internet Activity		Graeme Crallan	No Internet Activity	
Marcos Collier	No Internet Activity		Katoucha Niane	No Internet Activity	
Stephen D. Vance	No Internet Activity		Robert Nawojski	No Internet Activity	
K. Sivanesan	No Internet Activity		Dave Freeman	No Internet Activity	
Georgi Stoev	No Internet Activity		Florian Goebel	No Internet Activity	
Ahmed Emin	No Internet Activity		William Headline	No Internet Activity	