Exposing Inconsistent Web Search Results with Bobble

Xinyu Xing[1], Wei Meng[1], Dan Doozan[1],
Nick Feamster[1], Wenke Lee[1], and Alex C. Snoeren[2]

[1] Georgia Institute of Technology
[2] University of California, San Diego
{xxing8,wei,ddoozan3,feamster,wenke}@gatech.edu,
snoeren@cs.ucsd.edu

Abstract. Given their critical role as gateways to Web content, the search results a Web search engine provides to its users have an out-sized impact on the way each user views the Web. Previous studies have shown that popular Web search engines like Google employ sophisticated personalization engines that can occasionally provide dramatically inconsistent views of the Web to different users. Unfortunately, even if users are aware of this potential, it is not straightforward for them to determine the extent to which a particular set of search results differs from those returned to other users, nor the factors that contribute to this personalization. We present the design and implementation of Bobble, a Web browser extension that contemporaneously executes a user's Google search query from a variety of different world-wide vantage points under a range of different conditions, alerting the user to the extent of inconsistency present in the set of search results returned to them by Google. Using more than 75,000 real search queries issued by over 170 users during a nine-month period, we explore the frequency and nature of inconsistencies that arise in Google search queries. In contrast to previously published results, we find that 98% of all Google search results display some inconsistency, with a user's geographic location being the dominant factor influencing the nature of the inconsistency.

1 Introduction

Web search engines have emerged as the *de facto* gateway to the Internet, with the major players like Google and Bing locked in a heated battle to attract users from around the world. Personalization is a key tool for adding value to search results: Each search engine tailors search results not only to the query term, but also based on the profile of the user [1, 3]. Web search personalization aims to return the search results that are most relevant to each user, based upon the user's past search history, clicks, geographic location, device type, and other features that may help identify the user's preferences and predispositions [3]. Ideally, personalization identifies results that closely match the user's preferences and intent, improving user satisfaction and ultimately increasing revenue for the search engine.

In practice, Web search personalization may also hide certain results from users, when personalized results preempt search results that would have otherwise been included [7]. Because search personalization algorithms are effectively a "black box", users have little to no information about the information that personalization algorithms might prevent them from seeing. Moreover, personalization frequently occurs without the user's involvement—or even explicit agreement—so users may not even be aware that their search results have been tailored according to their profile and preferences. The goal of our work is to expose and characterize inconsistencies that result from personalization. In particular, we seek to quantify the extent to which search personalization algorithms return results that are inconsistent with those that would be returned to other users, and expose any differences to the user—in real time.

We present Bobble, a Chrome Web browser extension that allows users to see how the search results that Google returns to them differ from the results that are returned to other users. Bobble captures a user's search query and reissues it from a subset of over 300 world-wide vantage points, including both dedicated PlanetLab measurement nodes and the hosts of other consenting Bobble users. In contrast to research tools that have been developed to measure search personalization offline [5], we intend users to use Bobble while they browse the Web, providing them critical insight into how their online experience is being potentially distorted by personalization.

To understand the nature of the inconsistencies uncovered by Bobble, we study more than 75,000 real search queries issued by hundreds of Bobble users over nine months. We quantify the extent to which personalization affects search results and determine how users' Google search results vary based on factors ranging from their geographic locations to their past search histories. Our study study focuses exclusively on Google search, one of the more widely used search engines, but we expect that similar phenomena exist for other popular search engines. We find that 98% of Google Web searches return at least one set of inconsistent search results—typically from a vantage point in a different geographic region than the user, even though Bobble performs these searches without exposing any information that links to the searchers' Google profiles.

In sum, our study provides the first large-scale glimpse into the nature of inconsistent results that arise from search personalization and opens many avenues for future research. We quantify on how geography and search history may influence search results, but others have noted that many other factors (*e.g.*, device type, time of day) may also affect the results that a user sees for a given search term [5]. Bobble has been deployed and publicly available for 21 months; users and researchers can extend it to measure how other factors might induce inconsistencies in search results.

2 Related Work

Researchers have previously studied means to personalize Web search results. Dou *et al.* performed a large-scale evaluation and analysis of five personalized search algorithms using a twelve-day MSN query log [2]. They find that profile-based personalization algorithms are sometimes unstable. Teevan *et al.* conduct a user study to investigate the value of personalized Web search [11]. In contrast, we are less interested in the distinction between different personalization methods, and focus instead on the effects of a single search personalization algorithm. We aim to quantify the effects of different personalization factors on search inconsistency.

In a contemporaneous study, Hannak *et al.* measure the personalization of Google search. The bulk of their effort focuses on understanding the features leading to personalization, but they also conduct a limited study of real-world personalization by hiring 200 US-based workers to search a fixed set of 120 search terms using their own Google accounts [5]. They find that any given slot in the first page of search results has less than a 12% chance of being personalized. Directly comparing their result to ours is challenging, because we do not consider reordering. We instead focus on the set of results returned, not their order. Moreover, our study considers a larger set of real queries from a global set of locations, conducted over a longer time period. We find that almost all results are subject to some form of personalization. We do, however, replicate their method in Section 6 and find that personalization is more than twice as likely than their work suggests.

Personalization is not limited to Web search. Previous research has built distributed systems to understand the effect of information factors in a number of online services. For example, Mikians *et al.* develop a distributed system to demonstrate the existence of price discrimination on e-commerce sites and discover the effects of information factors on price discrimination [6]. They find the factors that contribute to price discrimination include the customer's geographic location, personal information, and origin URL. Guha *et al.* explore several approaches to determine how advertising networks adjust the advertisements that they display based on users' personal information [4].

3 Bobble

To identify inconsistencies in Google search results that result from personalization based upon geography or personal history, we design, implement and deploy Bobble, a distributed system that monitors and displays inconsistent search results that Google returns for user search queries in real time.

3.1 Design and implementation

Bobble has three components: a Chrome browser extension, hundreds of Chrome browser agents, and a centralized data collection server. Our Chrome browser extension[3] runs on a Google user's Chrome browser, and passively collects the Google user's searching activities including the Google user's search terms and corresponding search results. Chrome browser agents—running both inside users' Chrome browser extensions and in Chrome browser emulators that we install on PlanetLab nodes across the Internet—perform Google searches without signing in to a Google account or revealing a trackable browser cookie to Google. The central Bobble server coordinates the agents and archives users' search activities, their IP addresses, and the search results from the Chrome browser agents.

Bobble follows four steps to reveal inconsistencies in search results. When a user issues a Google search query (Step 1), Bobble browser extension delivers the search terms to the central Bobble server (Step 2), where they are placed in a global work queue. To protect user privacy, all subjects' Google identities are hashed by a one-way SHA-1 hash function. Asynchronously, Chrome browser agents periodically poll

[3] The Bobble Chrome browser extension is available from the Google Chrome store and our project website `http://bobble.gtisc.gatech.edu/`.

	with same browser	with Chrome agent	p-value
Windows	11 / 1,000	16 / 1,000	0.1725
Linux	23 / 1,000	21 / 1,000	0.7517
Mac	15 / 1,000	15 / 1,000	1.0

Table 1: The number of terms that generate inconsistent sets of search results when searching 1,000 distinct terms from Chrome browsers / agent on different OSes.

the Bobble server for pending search terms (Step 3) and reissue them locally as search queries to Google without signing into a Google account or revealing Google a trackable browser cookie (Step 4). Each agent pushes the results it receives from Google to the Bobble server.

To establish a baseline for comparing inconsistencies in search results, we would ideally like to also reissue the user's query locally from a separate browser session that is not signed into Google and does not pass session cookies to Google. We call these anonymous queries "organic", as they are as free as possible from user-specific influences (in contrast to queries that are issued when a user is logged in or passing browser cookies to Google). Unfortunately, collecting true organic results is challenging due to the technical and usability obstacles surrounding logging the user out in order to issue such a query from an extension running within the same Web browser. Instead, Bobble collects organic search results by issuing a duplicate query from a nearby Chrome browser agent. (Section 3.2 presents a detailed discussion of the effects of using a nearby agent to stand-in for the user's browser.)

3.2 Validation

To evaluate whether Bobble accurately reports results that regular users would actually receive, we first validate that Bobble's Chrome browser agent correctly emulates major version releases of Chrome browsers—specifically, that the results returned to a Bobble agent reflect those that would be returned to an actual query issued by a user in her Web browser. Second, we measure the effects of collecting organic search results indirectly by issuing queries from nearby agents as opposed to inside the user's browser.

Do Bobble agents emulate browser behavior? We begin by ensuring that the Google search results collected using the Chrome browser agent do not differ statistically from the results obtained when the query is issued from the Google home page viewed with the Chrome browser itself. We randomly select 1,000 unique search terms from the daily top-20 Google trending search terms between August 2011 and December 2011 and search each of these terms three times from machines running Linux, Windows, and Mac operating systems. On each machine, we run a Chrome browser agent and two Google Chrome browsers with the same release version. We use the Selenium Chrome driver [9] to automate the two Chrome browsers and one browser agent to perform the same Google search simultaneously.

One might expect that simultaneously issued queries from identical Web browsers would return identical sets of results, since the queries do not involve any search history and are issued from the same location at essentially the same time. While this expectation generally rings true, it is not always the case. Table 1 shows the number of

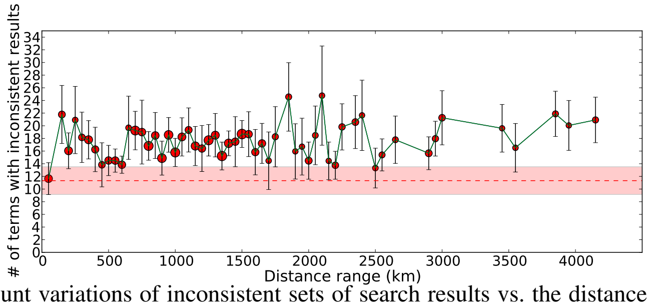

Fig. 1: The count variations of inconsistent sets of search results vs. the distance variations between a pair of PlanetLab nodes.

terms that generate inconsistent search results when comparing the first set of results returned to a Web browser to those returned to both the second instance of the browser and the Bobble agent; neither are non-zero. To test if the proportion of inconsistent results generated by our browser agent is statistically different from that of the browser, we conduct a two-sample proportion test. Table 1 shows that the proportion tests for the three operating systems are not statistically significant at the 0.05 level (*i.e.*, all p-values are greater than .05). In other words, we observe no significant difference observed in the proportion of inconsistent results generated by the Bobble browser agents and a real Chrome browser. We thus conclude that Bobble agents are reasonably accurate substitutes for real users executing search queries from within browser.

Are PlanetLab queries similar to real users? Bobble does not collect organic search results from within a user's own browser since this would require issuing duplicate queries from the user's browser and forcibly signing out the user and clearing the user's cookies. Instead, Bobble issues queries from an agent running on the closest PlanetLab node to obtain an approximation of what the Google user's organic search results would be. To identify how well this approximation holds with distance, we conduct the following experiment from 308 PlanetLab [8] nodes on which Bobble was deployed.

Using the same 1,000 search terms as before, Bobble browser agents search every term twice, back-to-back. Across the 308 nodes, 8–13 out of 1,000 terms generate inconsistent Google results with a 95% statistical confidence level[4]. This inconsistency may be due to caching, a sudden DNS change, updates to Google's indicies with their data center, or a myriad other possibilities. Regardless, we view this as a "noise floor" against which to judge inconsistency.

We now consider the number of terms that generate inconsistent search results when searches are performed on different PlanetLab nodes in the same country at varying geographic distances from each other. Figure 1 plots the average number of terms that result in inconsistent search results with a 95% confidence interval as a function of the distance between the two agents (according to Maxmind). The pink band represents the inconsistency observed from queries issued from the same node. Although there is no clear relationship between distance and consistency, only results returned to nodes

[4] When constructing confidence intervals, we consider searches from distinct browser agents to be independent trials from the same underlying distribution.

within 50 km of another node bear the same statistical level of resemblance as back-to-back queries issued by the same node. Hence, for the purposes of our study, we only consider queries where Bobble was able to collect organic search results from a PlanetLab node located within 50 km of the issuing agent. We selected the 50-km threshold because of the geographic distribution of PlanetLab nodes.

4 Data

On January 17, 2012, we released Bobble on both our project website and the Google Chrome store. As of October 25, 2012, we had collected 100,451 search queries. For each query, we record the corresponding Google search results returned to both the browser on which a Google user installs our Bobble Chrome extension and the Chrome browser agents that reissued the query. We obtain organic search results browser agents running on PlanetLab nodes no further than 50 kilometers from the user issuing the query. Using this criterion, we obtained organic search results for 76,307 of the search queries (75.96%).

To use 76,307 search queries for our analysis, we divided our data set into two categories: search queries issued by Google users while signed in to their Google accounts (signed-in Google users) and search queries issued by Google users while signed out (*i.e.*., anonymous Google users). There are 66,138 search queries (86.67%) issued by 174 distinct signed-in users, and 10,169 search queries (13.33%) issued by anonymous Google users.

5 Location-Based Inconsistency

We now analyze how geographic location affects search inconsistency. Search inconsistency contributed by geographic locations is a joint consequence of both location-based personalization and data diversity across different data centers. We analyze how geographic location contributes to search inconsistency that appears in different Google searches (Section 5.1) and validate that the inconsistencies we observe are in fact due to personalization, as opposed to inconsistencies across data centers (Section 5.2).

For each search query, we group the sets of search results from PlanetLab nodes into sets, each of which contains a unique result set. We compare the number of search results on the first page, as well as the rank, title and URL of each Google result. We use a nearby PlanetLab node's search results to represent the set of organic search results for a Google user in that region. If there is more than one unique search result set for a user's search query, we consider the results to be inconsistent, and we also deem geographic location to be a contributing factor to this inconsistency.

5.1 Results

We find that 74,594 out of the 76,307 search queries (97.76%) generate at least one inconsistent set of organic search results due to geographic location. Figure 2 shows the fraction of search queries that generate different numbers of inconsistent sets of

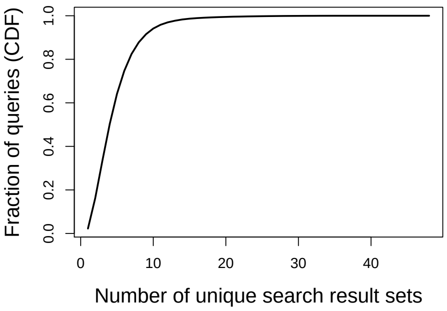

Fig. 2: CDF plot: the distribution of the number of search queries.

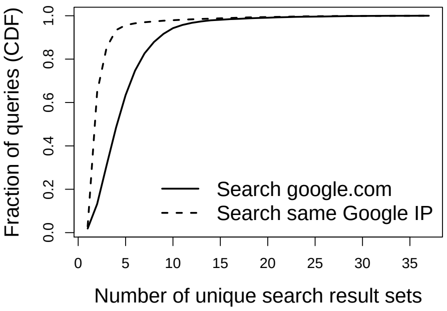

Fig. 3: The distribution of the number of search queries when sending queries to *google.com* and a Google IP address, respectively.

search results. This result indicates that organic search results of most Google search queries are tailored on the basis of the location where these searches are performed, even though Google users neither sign into their accounts nor uncover their browser cookies to Google personalized search services. In the following section, we further design a careful examination to explore whether the observed search inconsistency results from location-based personalization rather than data diversity across different Google data centers.

To quantify the effect of geographic location on search inconsistency, we classified the inconsistent search results in three ways:

- At least one search result appears in the top-three search results of other PlanetLab nodes but not at all in a Google user's organic search result set. We find that 23,394 out of 76,307 search queries (30.66%) give rise to this situation.
- At least one search results appears in the top-10 (but not top-3) search results of other PlanetLab nodes, but does not appear in a Google user's organic search result set; 65,939 out of 76,307 search queries (86.41%) fit this situation.
- At least one search result appears in the Google user's organic search result set but does not appear in search results of other PlanetLab nodes; 1,434 search queries out of 76,307 search queries (1.88%) fit this situation.

Considering the fact that the top-10 Google search results receive about 90% of clicks and the top-3 Google search results usually receive the most attention [10], the inconsistency that arises due to location likely has significant implications for a user's experience.

5.2 Distributed index inconsistencies

To validate the observed search inconsistency is in fact derived from location-based personalization rather than data diversity across different data centers, we conduct an experiment. In particular, we modify Bobble to attempt to isolate the inconsistency contributed by location-based personalization from that contributed by inconsistencies in the search index that may result from the index being stored across a globally distributed set of servers. We call these inconsistencies *distributed index inconsistencies*.

Experiment setup We direct the Chrome browser agents running on PlanetLab to send search queries not only to `google.com` but also to one particular Google IP address (74.125.130.100). Sending search queries to the same IP address can increase the likelihood that the search queries are processed by the same Google data center. Since the Chrome browser agents must perform any Google search twice (one on a particular data center and the other on a data center geographically nearby), which increases the risk of our Chrome browser agents being profiled as a search bot and challenged by Google CAPTCHA system, we limit our experiment to a subset of submitted daily search queries.

Quantifying distributed index inconsistencies We collect 23,362 search queries from 149 Google users. We then compare the numbers of unique search result sets for each collected search query when it is searched on `google.com` and the particular Google IP address. For all of the collected search queries, we observed that every search query sent to `google.com` nearly always generates a larger number of unique search result sets than it is sent to the particular Google IP address. Figure 3 shows that searching on `google.com` produces more inconsistent result sets than searching on the particular Google IP address does. This discrepancy likely results from the fact that directing a search to a particular Google IP address significantly reduces the influence of data diversity upon search inconsistency.

Another interesting observation from Figure 3 is that approximately 98% of search queries have at least one set of inconsistent search results, even though the influence of data diversity upon search inconsistency is nearly removed. Note Appendix indicates that the inconsistency within a single data center is minimal. We therefore believe that (1) these observed search inconsistency results from location-based personalization when the search terms are searched on the particular Google IP address, (2) location-based personalization contributes significantly to search inconsistency.

6 Profile-Based Inconsistency

We also explore how a user's profile (*i.e.*, search history) contributes to search inconsistencies. In particular, we treat the search queries (and corresponding results) independently based on the way that a user issues a search query. Table 2 summarizes our results. For the case of queries corresponding to signed-in users, 42,454 of 66,138 search queries (64.19%) generate results that are inconsistent with respect to the organic search results. For the anonymous users, 5,976 out of 10,169 search queries (58.77%) yield inconsistent search results.

In contrast to Hannak *et al.*'s prior study [5], we find that the profile-based personalization results in significant inconsistencies. Here, we replicate Hannak *et al.*'s experimental method. Figure 4 shows the percentage of search results changed at each rank in our data set. The average is 28.6%, compared to 11.7% as reported by Hannak

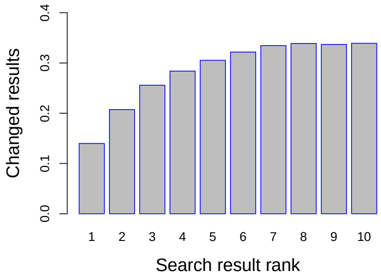

Fig. 4: % of search results changed at each rank.

Signed-in data set		Signed-out data set	
Location	Profile	Location	Profile
97.64%	64.19 %	97.80%	58.77%

Table 2: How location and user profile contribute to search inconsistency. Location has more effect on inconsistency than search history does.

et al. (see Figure 5 in previous work [5]). One possible reason for this discrepancy is the difference in the measurement method. Previous work recruited differnt Google users to search the same set of keywords, where the keywords were chosen such that they were deemed to not be related to user profiles. In contrast, we perform our study in a more natural setting because it measures the influence of the profile-based personalization using each user's own search queries. Because a user's past queries are typically relevant to personalization that may occur in the future, we observe that profile-based personalization has more influence on Google users' search results.

In addition to inconsistencies in the search result sets, we also discovered the following inconsistencies:

- For signed-in users, 22,405 out of 66,138 search queries (33.88%) have at least one search result that shows in the profile-based personalized search result set but not in the organic search result set.
- For anonymous users, 3,148 out of 10,169 search queries (30.96%) have at least one search result that shows in the profile-based personalized search result set but not in the organic search result set.
- For signed-in users, 7,352 out of 66,138 search queries (11.12%) have at least one search result that shows in the top 3 of the organic search result set but not in the profiled-based personalized search result set.
- For anonymous users, 1,484 out of 10,169 search queries (14.59%) have at least one search result that shows in the top 3 of organic search result but not in the profiled-based personalized search result set.

Table 2 also shows that the Google search inconsistencies resulting from signed-in users' profiles are stronger than those resulting from signed-out users' profiles. Finally, we also observe location-based factors introduce more inconsistencies than profile-based factors do.

7 Conclusion

We have designed, implemented, and deployed Bobble, a distributed system that tracks and monitors the inconsistency of search results for user search queries. Using Bobble, we collect user search terms and results and measure the search inconsistency that arise from both geographic location and search history. We find that the geographic location contributes more to search inconsistency than user search history, and that geographic location causes about 98% of search queries generate some level of search inconsistency. We have made Bobble publicly available to help users discover inconsistent results resulting from their own queries.

Acknowledgments. This work was partially supported by NSF Awards CNS-1059350, CNS-1162088, CNS-1255274 and a Google Focused Research Award.

References

1. More personalization on bing with adaptive search. `http://www.youtube.com/watch?v=CgrzhyHCnfw`.
2. Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.
3. Making search more relevant. `http://www.google.com/goodtoknow/data-on-google/more-relevant/`.
4. S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010.
5. A. Hannak, P. Sapieżyński, A. M. Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring Personalization of Web Search. In *Proceedings of the Twenty-Second International World Wide Web Conference (WWW'13)*, Rio de Janeiro, Brazil, May 2013.
6. J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM, 2012.
7. E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Press, 2011.
8. Planetlab: An open platform for developing, deploying, and accessing planetary-scale services. `http://planet-lab.org/`.
9. Selenium - web browser automation. `http://seleniumhq.org/`.
10. What is a #1 google ranking worth? `http://training.seobook.com/google-ranking-value`.
11. J. Teevan, S. T. Dumais, and E. Horvitz. Beyond the commons: Investigating the value of personalizing web search. In *Proceedings of the Workshop on New Technologies for Personalized Information Access*, 2005.

Appendix: Inconsistency within a Single Data Center

As a sanity check, we search the same set of 1,000 keywords in Section 3.2 by sending the corresponding queries twice in succession, but this time explicitly to the same Google IP address. We repeat the validation process sixteen times. Approximately 8 out of 1,000 (0.8%) keywords generate inconsistent search results on average, presumably because the Google indices stored on different servers in the same data center are different. We conclude that inconsistency within a single data center is minimal.