From Devices to People:
Attribution of Search Activity in Multi-User Settings

Ryen W. White[1], Ahmed Hassan[1], Adish Singla[2], and Eric Horvitz[1]

[1]Microsoft Research, Redmond, WA 98052 USA

[2]ETH Zurich, Universitätstrasse 6, 8092 Zürich, Switzerland

{ryenw,hassanam,horvitz}@microsoft.com, adish.singla@inf.ethz.ch

ABSTRACT

Online services rely on unique identifiers of machines to tailor offerings to their users. An implicit assumption is made that each machine identifier maps to an individual. However, shared machines are common, leading to interwoven search histories and noisy signals for applications such as personalized search and advertising. We present methods for attributing search activity to individual searchers. Using ground truth data for a sample of almost four million U.S. Web searchers—containing both machine identifiers *and* person identifiers—we show that over half of the machine identifiers comprise the queries of multiple people. We characterize variations in features of topic, time, and other aspects such as the complexity of the information sought per the number of searchers on a machine, and show significant differences in all measures. Based on these insights, we develop models to accurately estimate when multiple people contribute to the logs ascribed to a single machine identifier. We also develop models to cluster search behavior on a machine, allowing us to attribute historical data accurately and automatically assign new search activity to the correct searcher. The findings have implications for the design of applications such as personalized search and advertising that rely heavily on machine identifiers to custom-tailor their services.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; *selection process*.

Keywords

Search activity attribution; Multi-user settings.

1. INTRODUCTION

User identifiers are central to a range of applications on the Web, including behavioral analysis [40], personalized search [36], and online advertising [7]. These *machine identifiers* are assigned to the machine via mechanisms such as browser cookies or toolbars. With single identifiers tied to a machine, applications and services operate under the implicit assumption that identifiers refer to users. However, for shared machines in homes and workplaces this is often an erroneous assumption. Although recent estimates suggest that 75% of U.S. households have a computer, in most households machines are shared between multiple people [20]. Different people may use the shared machine at different times, but to a remote observer all activity is associated with a single identifier, and people's search behaviors will be intertwined in search logs. This creates a noisy behavioral signal, and importantly, a challenge for analyzing search behavior, especially long-term behavior that has utility in many applications, such as search personalization [37].

Let us consider some real-world data gathered from a panel of millions of Web searchers recruited by comScore (comscore.com), an
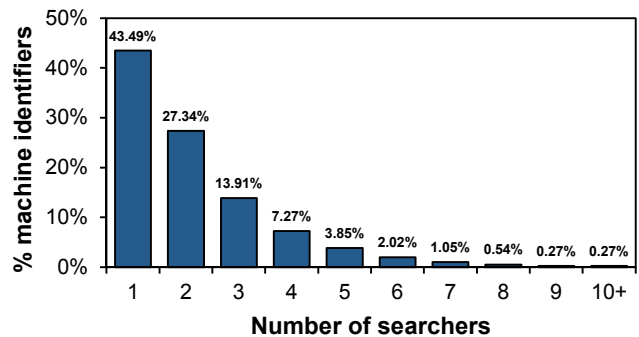
Figure 1. Percentages of machine identifiers in our dataset comprised of each number of searchers (k), from 1 to 10+.

Internet analytics company. In addition to a machine identifier, similar to that obtainable via Web browser cookies and other applications, panelists have a *person identifier* and are required to sign-in prior to use to indicate that they are searching on the machine at a particular time. Since we have both machine identifiers and person identifiers, we can compute the frequency with which multiple people are observed searching on a particular machine, as well as other characterizations of searching and searcher interests reported later. We can also use these data as ground truth in developing models to estimate the number of searchers within a machine identifier, and in attributing search activity observed historically to specific searchers. Figure 1 shows the fraction of machine identifiers in the dataset that are comprised of different numbers of searchers.

It is striking from the figure that 56.5% of machine identifiers comprise the search activity of more than one person. Although we report statistics from only one data source, the data are purposely gathered from a representative sample of United States households [21]. The mean and median number of searchers per machine observed in the data are 2.39 and 2 respectively, aligning well with U.S. census estimates on the size of households (mean=2.55, census.gov/hhes/families). We note that the comScore data used in our study are not proprietary; other researchers can purchase the logs from comScore, and can replicate and extend our findings.

We envisage that the performance of personalization and ad-matching would likely be enhanced if user-centric signals and analyses were used. We also see privacy benefits of being able to accurately segregate searcher activity within a machine. Providing a means to preventing the unintended sharing of sensitive information between the searchers on the same machine and help ensure that only necessary information is shared with search providers [25][26]. Despite the importance of accurately attributing search activity, to our knowledge, we know of no prior research on this topic beyond the level of machine identifiers. We address this shortcoming in this paper by characterizing variations in search behaviors within machines and developing predictive models to assign observed search activity to the correct *individual*. In doing so, we can capitalize on well-documented aspects of human behavior such as the bursty nature in which human events typically occur [4].

We make the following research contributions with our work:

- Introduce the challenge of *search activity attribution* and through empirical analysis demonstrate its potential significance for research and practice in Web search engines;
- Characterize key variations in behavioral, temporal, and topical signals associated with differences in the number of individual searchers associated with a particular machine;
- Develop models to accurately predict the presence of multiple searchers associated with a machine identifier using a range of signals, and using regression to quantify the likely number of searchers associated with an identifier. This inference alone could help decide which identifiers contain noise, inform decisions about when long-term histories can be trusted, and when more computationally-expensive methods, such as clustering of search histories, need to be applied, and;
- Leverage the estimated number of searchers from the previous step, we cluster search activity based on a range of similar features and show that we can accurately assign new search activity to the correct searcher. The estimated searcher count from the regression alone is insufficient since we do not have a representation of those searchers' activity, and need such a representation to handle the assignment of new queries.

We focus on the Web search domain given the nature of the data available and its importance, but the activity attribution challenge applies to a number of other domains including online advertising, audio signal processing, and fraud detection.

The remainder of this paper is structured as follows. Section 2 describes related work in areas such as behavioral analysis and personalization. Section 3 provides an overview of the data used in the study. Section 4 presents a characterization of the data. Section 5 describes the prediction of whether multiple searchers comprise a machine identifier (classification task) and estimating the number of searchers that contribute to the search activity associated with a machine identifier (regression task). Section 6 uses the output of the regression and through clustering, addresses the challenge of assigning activity from a particular machine identifier to the correct searcher. We discuss our findings and their implications for the design of online services in Section 7, and conclude in Section 8.

2. RELATED WORK
Related work on this topic falls into a few key areas: (1) log-based analysis of search activity, (2) individual differences in search behavior (which could help differentiate searchers in logs), (3) personalization of Web search and advertising (capitalizing on differences in searcher interests), and (4) related application domains (namely Website access analysis, fraud detection, and blind signal separation). We discuss each of these research areas in turn.

Logs of search behavior have been analyzed extensively in the Web search and data mining communities to better understand how people search [40], predict their next online actions [16][27], predict their future interests [18], improve search engines [23][35], and understand in-world activities from long-term search log data [30]. The longer-term analysis of behavior in particular has leveraged machine identifiers assigned based on Web browser cookies or add-ins to attribute actions from a single identifier to the same searcher to study variations in behavior and interests over time [30][35][40]. However, as shown in Figure 1, the machine identifiers used in log analysis may not be always reflect a single searcher's behavior.

Information scientists have studied individual differences in search strategies, tactics, and performance, and other factors such as cognitive styles and domain expertise [2][6][31][38] that can influence search behavior and task outcomes. These studies provided detailed

modeling of search behavior, often coupled with surveys to understand motivations, but have small numbers of searchers and tasks.

Individual behavioral differences may help distinguish searchers. Large-scale log analyses examined the relationship between search and domain expertise and behavior [39][42]. White et al. [42] found that domain experts are more successful than novices (when searching in the domain of their expertise) and achieve this success via different vocabulary, sites, and broader search strategies. White and Drucker [40] identified *navigators* (consistent search and browsing patterns) and *explorers* (varied search and browsing patterns). Recent studies have examined differences in how people inspect result pages via eye-gaze tracking [17] and mouse cursor tracking [8].

Search preferences are personal and research on personalizing retrieval [33][36] has found that implicitly-gathered information such as browser history, query history, and desktop information, can be used to improve result relevance. These methods rely on accurate attribution of search activity to individual searchers. Short-term behavior from within the current search session has been used for result ranking [44] or predicting future search interests [41][43]. Teevan et al. [36] showed that personalization improves as more data was available about the current searcher. Long-term behavior has been used to personalize search [34], including using previous queries associated with the pursuit of similar information needs [35]. Models can use different sources, ranging from specific query-URL pairs which have high precision but low coverage [37] to more general methods using topical representations of searcher interests [29][34]. Similar methods have been applied in the advertising domain, where behavioral data associated with an identifier are used to tailor advertisements accordingly [11][45]. Rather than storing an individual's profile information on the server, other methods propos personalized advertising using client-side storage of profile information to address privacy concerns [7]. In all of these applications, models use long-term search behaviors for identifiers assumed to be associated with the same individual. However, searcher interests can differ greatly between individuals and failing to consider these variations can lead to noisy models and sub-optimal personalization performance. Research has studied the reliability of user identifiers but primarily tracking users across identifiers (ameliorating cookie churn) [14] and not on the segregation of observed search activity *within a machine identifier* as we target here.

Our research shares challenges with other research areas, such as clustering Web site visits, fraud detection, and signal processing. Website designers are interested in the interests and intentions of those who visit their sites. Cadez et al. [9] proposed model-based clustering of visits by topical interest. Moving beyond information-seeking, the goal in fraud detection is to identify suspicious changes in a person's behavior, where observed activity may not be representative of their typical actions. This involves building a profile over time and looking for anomalous behaviors [19], a goal that does not match with the objectives of our work. In signal processing, blind signal separation (BSS) [1][10] (and instantiations such as independent component analysis [13]) involves separating source signals from observed mixtures, typically the output of an array of sensors. BSS been successfully applied in domains such as communications [3] and medicine [28]. However, the applicability of these methods to our scenario is less clear. We employ methods widely adopted in the search and data mining communities.

The methods presented in this paper extends previous work in a number of ways. The search activity attribution challenge is an important problem for search providers that, to our knowledge, has not been addressed previously. Second, we present a detailed characterization of within-machine variance in search behavior that

Table 1. General statistics of the dataset used in our study.

Statistic	Value
Total number of queries	576,470,390
Total number of machines	1,748,425
Total number of searchers	3,836,037
Average queries / machine	328.89 (stdev=1279.80)
Average duration (in days) / machine	126.07 (stdev=171.29)

Figure 2. Percentage increase in search activity from multi-searcher machines vs. single searcher machines.

motivates the development of predictive models. Third, we develop models to attribute search activity to people and tackle the within-device segmentation and assignment challenge. Finally, rather than searching for anomalous activity or isolating source signals as in related research, we focus on the accurate attribution of new search activity to a particular person and discuss the implications of this for Web search, advertising, and other online services.

3. DATASET

The data that we used for our study was provided under contract by the Internet analytics company comScore. comScore recruited an opt-in consumer panel that has been validated to be representative of the online population and projectable to the United States population [21]. Millions of panelists provide comScore with explicit permission to passively measure all of their online activities using monitoring software installed on their computers. In exchange for joining the panel and providing search data, participants are offered a variety of benefits, including computer security software, Internet data storage, virus scanning, and chances to win cash or prizes.

The data comprised unfiltered search queries on major Web search engines such as Google, Bing, and Yahoo, collected over a two-year period from mid-2011 to mid-2013. The logs contained the text of queries, search result clicks, and the time that the events occurred (in searcher's local time). Importantly for our study, the logs also contained a machine identifier (assigned to the machine) and a person identifier (assigned to each person who used the machine). An application is installed on the machine to record search activity and searchers are required to indicate to the logging software that they are searching at any given time. Machine-based identifiers are used in a range of online applications, either through Web browser cookies or other mechanisms such as search-provider toolbars; so their use in this study reflects reality. To remove variability caused by cultural and linguistic variation in search behavior, we only include log entries from the English-speaking United States locale. Figure 1 in the previous section summarizes the number of people associated with each machine identifier. Table 1 shows some basic data statistics including the average duration in days, defined as the time between the first and last observed query on each machine.

Using the data described above, we can compute and examine behavioral and temporal features. However, there are other features that may vary based on the number of searchers on a machine, including the topic of the content viewed and the complexity of that information (e.g., perhaps reflecting age differences in searchers – suggesting the presence of more than one individual). To enable a richer analysis and of different feature sets we employed classifiers to assign topical labels to the clicks using the hierarchy from the Open Directory Project (ODP, dmoz.org) [5] and the complexity of the queries/results, based on estimates of their U.S. school grade level (on a 1-12 scale) [12]. We describe the behavioral, topical, temporal, and other features in more detail later in the paper.

4. MULTI-USER SEARCH BEHAVIOR

In this section, we seek to understand if and how search behavior ascribed to a single machine identifier changes with the number of searchers associated with that identifier (available via the person
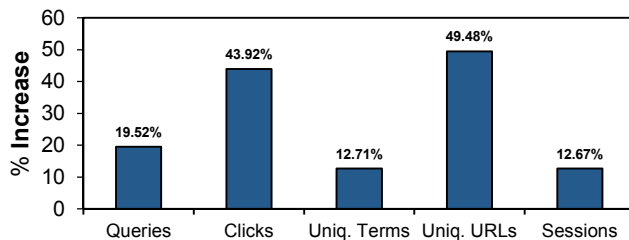
identifier in the comScore logs). More specifically, we examine several characteristics of machines with different number of searchers focusing on behavioral, temporal, topical, and content dimensions. Our features are computed per day or per week (depending on the feature) to reduce the effects of differences in the search history length. We begin with the behavioral characteristics.

4.1 Behavioral Characeristics

At the outset of our analysis, we examined a number of different measures characterizing the search activity of searchers on single- and multi-searcher machines. In particular, we examine: (1) the number of queries per day, (2) the number of clicks per day, (3) the number of unique query terms per day, (4) the number of unique clicked URLs per day, and (5) the number of search sessions per day. To calculate the number of query terms, we convert all queries to lowercase, replace contiguous whitespace with a single space, and segment the query into terms using space as a separator. To segment queries into sessions, we introduce a session break if the searcher was idle for more than 30 minutes. Similar criteria have been used to demarcate search sessions, e.g., [16][40].

We notice that the average search activity from multi-searcher machines is significantly ($p < 0.01$) larger than the average search activity from single-searcher machines, across all measures studied in the rest of this section, including temporal and topical-content features. We report percentage gain for the different measures in Figure 2. All differences we report in this section are statistically significant at $p < 0.01$ using two-tailed t-tests unless otherwise stated.

4.2 Temporal Characteristics

In addition to exploring properties of search activity on single- and multi-searcher machines, we also examine the temporal usage behavioral patterns as the number of searchers per machine increases.

Day Entropy: One interesting characteristic of the temporal behavior patterns may be the distribution of search queries across days. We might expect multi-searcher machines to have search activity that is more disparate across days (given different time constraints from work, schooling, etc.). To validate this hypothesis, we divided the observed queries into seven buckets corresponding to days of the week. We then compute the normalized entropy of the query distribution across days as:

$$H = \frac{-\sum_{i=1}^{n} p(x_i) \log(p(x_i))}{\log(n)} \quad (1)$$

where n is the total number of outcomes (the seven days). A value of zero would suggest that there is no uncertainty in the daily distribution of queries (i.e., all queries occur on the same day of the week). While a value of one would suggest maximum uncertainty (i.e., queries are evenly distributed across all seven days).

The day entropy of the identifiers is shown in Figure 3 for machines with 1–5 searchers using a box-and-whisker plot. The horizontal segments inside the boxes represent the median entropy, the top and
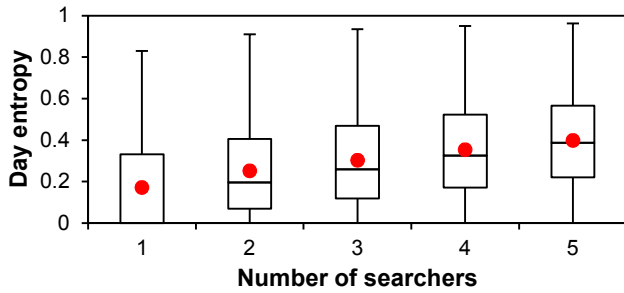
Figure 3. Box-and-whisker plot for day entropy for machines with diff. # searchers. Mean is dot. Median is horizontal line.
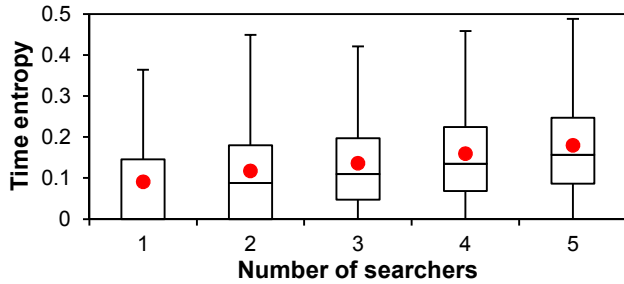

Figure 5. Box-and-whisker plot for topic entropy for machines with diff. # searchers. Mean is dot. Median is horizontal line.

Figure 4. Box-and-whisker plot for time entropy for machines with diff. # searchers. Mean is dot. Median is horizontal line.

Figure 6. Box-and-whisker plot for readability entropy with diff. # searchers. Mean is dot. Median is horizontal line.

bottom of the box denotes the first and third quartiles, and the whiskers denote the maximum and minimum. We also added the mean as a circle. For simplicity, we ignored machines with more than five searchers since they account for less than 5% of the machines (see Figure 1). Figure 3 shows that as the number of searchers per machine grows, the entropy increases suggesting that machines with multiple searchers have more diverse daily usage patterns and that this diversity grows as the searcher count increases.

Time Entropy: We now consider another aspect of the temporal usage patterns concerning the *time of the day* at which the search activity occurs. We divide search queries into six different equally-sized time buckets corresponding to following time ranges: morning (6am-10am), midday (10am-2pm), afternoon (2pm-6pm), evening (6pm-10pm), late night (10pm-2am) and overnight (2am-6am). We compute the normalized entropy of the time buckets as described earlier. The results are shown in Figure 4. In a similar way to day entropy, we can observe is a clear trend of entropy increasing as the number of searchers increases. One explanation is that searchers may have fixed time preference for when they search.

4.3 Topic and Content Characteristics

Topical and content information has been used extensively to model search behavior and to capture intent [34][43]. We wanted to understand the relationship between topic/content and the number of searchers per machine. This may help us understand whether topical profiles of single-searcher machines differ from those of multi-searcher machines (which could be useful for the prediction tasks described later). We examine three different aspects of topics and the nature of the content that searchers seek: (1) topic entropy, (2) readability level entropy, and (3) between-topic associations.

Topic Entropy: We assigned a topic to each search query based on the plurality label of the topics assigned to its clicked URLs in historic log data. To do this, we used the content-based classifier described and evaluated in [5]. The classifier assigned ODP category labels to URLs. ODP is an open Web directory maintained by a community of volunteer editors. It uses a hierarchical scheme for organizing URLs into categories and subcategories. Many previous
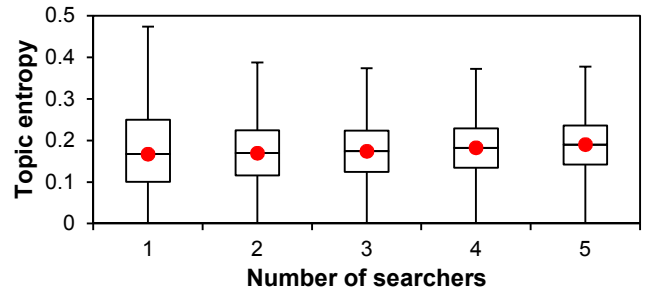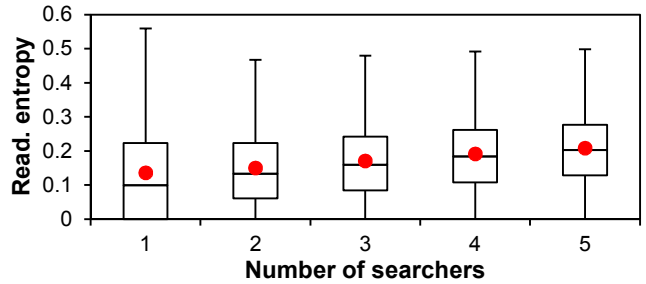
studies of Web search behavior have used ODP to assign topics to URLs, e.g., [32]. Queries that received no clicks were ignored.

After assigning topics to queries, we calculate topic entropy as we described previously and we show the results in Figure 5. The figure shows that topic diversity increases as the number of searchers increases even though the difference are smaller than the variations in the temporal entropies, the differences between the means are still statistically significant at $p < 0.05$ using a two-tailed t-test.

Readability Entropy: Another aspect that may be correlated with topicality is the readability level of the text of the queries. We might expect that the population of searchers sharing the same machine have different ages and this affects the sophistication of their search queries. We are likely to see low variance in readability level in single-searcher machines, but on multi-searcher machines the variance may be high, especially if the searchers are of different ages.

Previous work has studied the problem of automatically assigning readability level to text [12]. The readability level of any text fragment can be assessed by assigning a value on a 12-point scale corresponding to U.S. school grade levels. The reading level predictor adopts a language modeling approach using a multinomial Naïve Bayes classifier [12]. The entropy over the reading levels for machines with different number of searchers is shown in Figure 6. The figure shows that the variance in readability level clearly increases with searcher count suggesting it could help predict that count. We show in our findings later in the paper that this is indeed the case.

Topic Association: Many search applications such as personalization, query suggestion, targeted advertising are interested in answering the following question: If a person searches for topic *A*, are they also likely to be interested in topic *B*? In order to understand whether there is a difference between such associations on single-searcher machines and multi-searcher machines, we conducted the following experiment. We derived the association between all pairs of topics using queries from single-searcher machines. We refer to these the "true" associations. We also derive them from machines with multiple searchers assuming that behavior is not dependent on
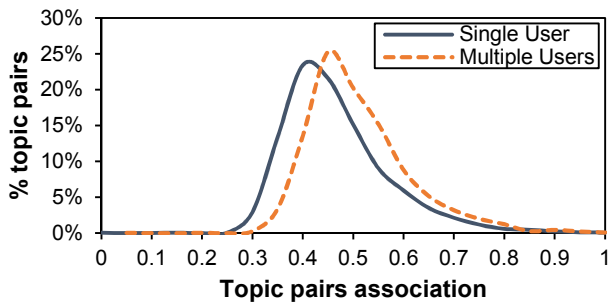
Figure 7. Distribution of topic pairs association from single-searcher machines (true dist.) and multi-searcher machines.
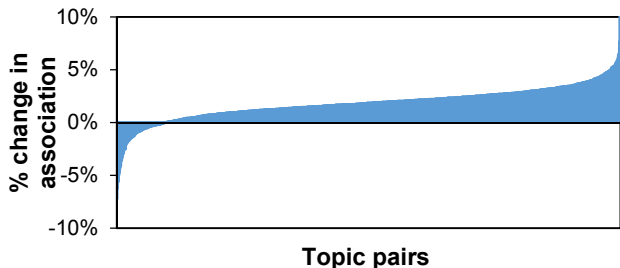

Figure 8. % change (error) in topic association for multi-searcher vs. single-searcher (truth). Positive change shows multi-searcher machines overestimate truth for 90% of pairs.

the searcher count. If this assumption were correct, we would expect to observe no difference between the two ways of computing the associations. If that assumption is incorrect (i.e., we observe co-occurring topics on multi-searcher machines that do not typically occur on single searcher machines), then we would expect to see differences between the outcomes of the two methods for computing associations.

To compute topic associations, we assume that two topics T_i and T_j co-occur if we observe two queries Q_i and Q_j in the same *time bucket* (we use the same time buckets defined earlier in this section) such that the topic of Q_i is T_i and the topic of Q_j is T_j. Given this co-occurrence definition, we assess topic association by computing the normalized point wise mutual information between them:

$$NPMI(T_i, T_j) = -\log \frac{p(T_i, T_j)}{p(T_i)p(T_j)} \Big/ -\log p(T_i, T_j) \qquad (2)$$

To compare topic associations, we plot the distribution of all topic associations derived from single searcher machines versus multi-searcher machines in Figure 7. We see from the figure that the associations from multi-searcher machines are over-estimating the true associations. This is also shown in Figure 8, which shows the percentage change in association between all topic pairs when computed from multi-searcher machines as compared the true topic associations (90% of the pairs had a positive change). These findings show that when we observe topics co-occurring on the same machine that typically have low association, we may be able to reliably estimate that there are in fact multiple searchers on that machine.

4.4 Summary

In this section, we have characterized some key aspects of search behavior in single- and multi-searcher machines. We have shown that there are significant differences in terms of search activity volume and temporal usage patterns. We have also shown that, although there are limited differences in topical variance across machines with different searchers, there are significant differences in terms of readability level (which may provide insight on searcher

age) and topical associations. In the next sections, we describe models that leverage similar features to estimate the number of searchers per machine, cluster search behavior on a machine, and attribute historical search activity and correctly assign new activity.

5. PREDICTING MULTI-USER SEARCH

Important tasks in the attribution scenario are to be able to (a) estimate whether a machine identifier comprises multiple searchers (binary classification), and (b) estimate the *number* of searchers whose search activity comprises that identifier (regression). Using features such as those described in Section 4 we developed predictive models to perform these tasks. The true number of searchers on a machine (k) is available from the comScore data. The estimated number of searchers on a machine can be used to guide the application of clustering methods, such as k-means clustering [28], that we employ in Section 6. The binary classification task is less challenging than predicting the number of searchers, but could still have utility for a search engine. For example, the inference can help gate the application of more sophisticated, and computationally expensive, analysis of the search history from a particular machine identifier (such as clustering), or it can help decide whether personalization methods should be employed for an identifier even without clustering. In this section, we report the performance on each task, beginning with the experimental setting and features used.

5.1 Experimental Setting

Using the data described Section 3, we extracted features of the historic behavior from each of the 1.75M machine identifiers in a similar way to the characterization in the previous section. The specific features are described in the next subsection, but they were informed by the characterizations described thus far. As before, since there were varying history lengths in our data, and in longer histories we may be more likely to observe multiple searchers, we computed all features normalized *per day* or the case of the few weekly features (e.g., day entropy as in Figure 3) normalized *per week*. The presence of variable history lengths closely resembles how the classifier would likely be used in practice: at some point it would be applied to historic logs, containing potentially different amounts of search history for each machine identifier in our logs.

We used Multiple Additive Regression Trees (MART) [23] for both the classification and regression tasks. MART uses gradient tree boosting methods for regression and classification. Advantages of employing MART include model interpretability (e.g., a ranked list of important features is generated), facility for rapid training and testing, and robustness against noisy labels and missing values. We also experimented with filters on the total number of days required to include identifiers in the analysis (e.g., filtering to only those identifiers with at least seven days between the first and last observed query), and they had little effect on the performance of the predictive models learned. Therefore, we did not use such filters so as cover all identifiers and remove the need for a threshold.

Ten-fold cross validation was employed across ten experimental runs and the performance numbers are reported as averages across those runs. Since the unit in the experiments was machine identifier and predictions are made at the identifier level, during the experiment an identifier was either in training set or the testing set, but not both. Evaluating *between* machine identifiers in this way improves the robustness and generalizability of our findings since the predictors could therefore scale to new, unseen machine identifiers.

5.2 Features

We devised around 70 features of the search behavior observed from the machine identifier. Table 2 presents a description of the specific features from each of the five classes that are used in our

prediction experiments. Many of these were informed directly by the characterization performed in the previous section. The five classes and their features are summarized as follows.

Temporal: These features describe the time at which the query is issued, in terms of time of day (in four-hour time *buckets*) and day of week, as well as variations in these features per number of unique time ranges and the entropy of those distributions defined as earlier. The rationale behind including the temporal features is that people may be likely to only access the machine at certain times given time constraints from commitments such as employment and schooling, as well as sleeping and eating. We include the entropy of searching for adult material since that is a sensitive subject area and unconstrained searching for pornography may be suggestive of fewer temporal limitations. We also include the timespan between queries and sessions as features, with the expectation that search activity may be sparser given fewer searchers associated with the machine.

Topical: Since searchers may have different topical interests, we encoded a number of aspects of the topicality of the queries issued and the results selected. To do this, we employed two classifiers: (1) a content-based classifier that assigns topical categories from the top-two levels of the ODP (e.g., "*Arts/Television*") to URLs as described earlier and in [5], and (2) a proprietary classier provided by the Microsoft Bing search engine that assigns topical categories to queries (e.g., images, movies, health); these are referred to hereafter as *query categories*. We compute a number of measures of the variation in the topic for the clicked URLs for the full category label, the top level only, and the first category accessed in the session, since that may more accurately reflect searcher intentions [9]. We also featurize changes in topics during query transitions to help represent the dynamics of the search interests. Finally, we focus on a number of specific classes which may be indicative of the number residents in the household and the number of searchers on the machine (e.g., the fraction of queries about shopping for child products (suggesting more searchers) or the fraction of queries on nightlife (suggesting fewer searchers)). We also compute the average distance (steps) in ODP, between pairs of topics accessed by searchers in sequence featurizing aspects of the topical focus. Finally, we featurize the topical association that is shown as useful in Section 4.3.

Behavioral: Motivated by the results of the behavioral characterization presented earlier in the paper (Figure 2), this captures aspects of the search behavior on the machine, and includes features such as the number of sessions, the number of queries, and average query length. The rationale is that significant variations in search behavior on machine may be attributable to multiple searchers. We also capture variations in the average click rank and the entropy of the clicks (how diverse those are on average, as in [15]). We also include the average historic frequency of queries from the Microsoft Bing search engine query logs (from a time period preceding the comScore logs used in this study) since this may provide insight into the nature of searchers' information needs independent of query topic (e.g., less popular queries suggest specific needs).

Content: These features capture variations in the nature of the information that searchers of the machine seek and access. The rationale is that with more searchers on a machine, there is likely to be more variation in the types of content accessed. This class includes information on the resources visited (URLs and Web domains), and top-level domains such as .com and .org, shown to reflect searcher differences (in expertise) in previous work [41]. Note that this includes the readability level estimates for both the queries and the pages visited (using the classifier described in [12]), and variations in those estimates across queries and clicks in the search history. If there were multiple searchers, especially a mixture of

Table 2. Prediction features. "P" denotes time-of-day class.

Feature	Feature Description
Temporal class	
FractionWeekday	% of queries on a weekday
FractionWeekend	% of queries on a weekend
FractionQueries_Morning^P	% of queries at 6am–10am
FractionQueries_Midday^P	% of queries at 10am–2pm
FractionQueries_Afternoon^P	% of queries at 2pm–6pm
FractionQueries_Evening^P	% of queries at 6pm–10pm
FractionQueries_LateNight^P	% of queries at 10pm–2am
FractionQueries_Overnight^P	% of queries at 2am–6am
NumTimeBuckets^P	# of time buckets per day
NumDays	# of days per week
TimeEntropy^P	H(Time bucket distribution) per day
DayEntropy	H(Day bucket distribution) per week
TimeBetweenQueriesAverage	Average time between queries
TimeBetweenQueriesVariance	Variance in time between queries
TimeBetweenSessionsAverage	Average time between sessions
TimeBetweenSessionsVariance	Variance in time between sessions
AdultTimeEntropy	H(Adult time bucket distribution) per day
Topical class	
TopicEntropy	H(ODP category assigned to clicks)
FirstTopicEntropy	First ODP category in session entropy
TopTopicEntropy	Top-level ODP category entropy
QueryCategoryEntropy	H(Query category)
FractionUniqueTopics	% of ODP topics unique
FractionUniqueQueryCategories	% of query categories unique
TopicDistance	Average inter-topic distance in ODP
TopicDistanceVariance	Variance inter-topic distance in ODP
NumUniqueTopics	# of unique ODP categories
NumUniqueQueryCategories	# of unique query categories
FractionTransitionsTopicShift	% of query transitions with ODP change
NumUniqueTopLevelTopics	# unique top-level ODP categories
FractionUniqueTopLevelTopics	% unique top-level ODP categories
FractionQueries_Adult	% queries on Adult query category
FractionQueries_Cooking	% queries on ODP "Cooking"
FractionQueries_Family	% queries on ODP "Family"
FractionQueries_KidsAndTeens	% queries on ODP "Kids & Teens"
FractionQueries_Nightlife	% queries on ODP "Nightlife"
FractionQueries_ShoppingChild	% queries on ODP "Shopping/Children"
FractionQueries_VideoGames	% queries on ODP "Video Games"
TopicAssociation	Average topic association per day
Behavioral class	
NumSessions	# search sessions
NumQueries	# search engine queries
NumUniqueQueries	# unique queries
NumUniqueQueryTerms	# unique query terms
FractionUniqueQueries	% query terms that are unique
QueryLength	Average query length (in characters)
QueryLengthVariance	Variance in query length
NumClicks	# result clicks
AvgClickRank	Average rank position of result clicks
ClickEntropy	H(Search result clicks), defined as in [15]
HistoricQueryPopularity	Historical query popularity in Bing logs
Content class	
DomainEntropy	H(Web domains visited), from clicks
QueryReadingLevel	Average query reading level (1-12) [12]
QueryReadingLevelVariance	Variance in query reading level
QueryReadingLevelEntropy	H(Query reading level)
PageReadingLevel	Average landing-page reading level
PageReadingLevelVariance	Variance in landing-page reading level
PageReadingLevelEntropy	H(Landing page reading level)
NumUniqueTopLevelDomain	# unique top-level domains (e.g., .com)
FractionUniqueDomains	% of unique top-level domains
FractionUniqueURLs	% of unique URLs
NumUniqueURLs	# of unique URLs
NumUniqueDomains	# of unique Web domains
FractionUniqueQueryURLs	% unique query-URL pairs
Referential class	
FractionReferenceFamily	% queries containing term "family"
FractionReferenceHousemate	% queries with reference to housemate

Table 3. Classification performance for each classifier, ordered by classification accuracy. All differences significant using *t*-tests at *p* < 0.001 for accuracy and AUC for each classifier versus marginal and versus *All*.

Features	Accuracy	Pos. Prec.	Pos. Recall	Neg. Prec.	Neg. Recall	AUC
All	0.8635	0.8662	0.8973	0.8597	0.8196	0.9366
Temporal	0.8552	0.8531	0.8986	0.8582	0.7986	0.9267
Topical	0.8324	0.8399	0.8694	0.8218	0.7824	0.9105
Content	0.8271	0.8351	0.8651	0.8157	0.7776	0.9055
Behavioral	0.8096	0.8027	0.8795	0.8208	0.7185	0.8827
Referential	0.6450	0.8751	0.4342	0.5552	0.9193	0.6871
Marginal	0.5651	0.5651	1.0000	0.0000	0.0000	0.5000

adults and children, then we would expect to observe variations in the reading level. There were some indications of this in Figure 6.

Referential: The last class of features that we considered involved references to other people, specifically the use of the word "family" or people who were likely to share accommodation with the current searcher (e.g., husband, child, roommate, spouse).

5.3 Prediction Results

We now present the findings of our experiments on both of the prediction tasks, beginning with the classification results.

5.3.1 Classification

The classification task involves the binary prediction of whether a machine identifier comprises multiple searchers. For the baseline in this task we use the marginal, which assumes that we always predict that a machine identifier is composed of multiple searchers (given the distribution reported in Figure 1). We report on the average accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC) across all experimental runs. Table 3 shows the performance measures for the full model (labeled as "All"), each of the five classes, and the marginal baseline.

The table shows that the performance of the classifier trained on all features was strong. The accuracy was 0.8635 and the AUC was 0.9366, significantly higher than the marginal classifier according using two-tailed *t*-tests (*p* < 0.001) paired on the folds. Table 3 also shows the performance of classifiers for each of the five feature classes separately. The table shows that many of the feature classes perform well in isolation (all classifiers significantly outperformed the marginal at *p* < 0.001), although not as well as the complete combination of features in the *All* model. These findings concur with the characterizations presented in Section 4, which clearly showed that there are many ways in which multi-searcher search activity within a single machine identifier can be detected.

Figure 9 shows the ROC for the top-performing *All* model and, for reference, a single point denoting the performance of the marginal classifier. Since some of our features rely on search providers running sophisticated classifiers on queries and visited content to compute measures such as topicality or reading level, it is important to quantify prediction performance with only a minimal set of features. Since the Temporal class performed particularly well in the analysis presented above, were also curious about how well we could perform for a pruned set of temporal features where we only featurize the time of the day on which search was performed on the machine. The eight features used for this pruned time-of-day only model are marked in Table 2 with superscript "P." We show that this model performs well, with accuracy of 0.8272 (AUC=0.8953), even though only a small subset of our features were used. To understand the performance of this classifier across the range of its discrimination threshold, we also plot its ROC curve in Figure 9.
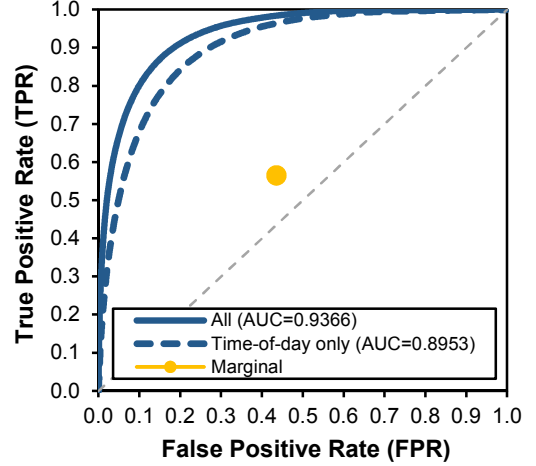

Figure 9. ROC curve for all features versus time-of-day only. Marginal is also shown at TPR=0.5651, FPR=0.4349.

The most useful features in the pruned temporal classifier were *TimeEntropy* (the variation in the times of day at which searches are performed on the machine) and *NumTimeBuckets* (the number of distinct four-hour time windows with search activity). Both features are likely to capture the dispersion of queries across the course of a day; which may be indicative of multiple searchers, especially given that people cannot search on the machine simultaneously.

5.3.2 Regression

In addition to predicting whether an identifier is composed of multiple searchers, we can also estimate the *number* of searchers comprising the logs of that identifier (the *k*-value described earlier). We focus on *k* in the range [1,10], folding the rare cases where $k > 10$ (0.05% of the data) into the $k = 10$ bucket. We frame this problem as a regression task and once again use MART, although this time to perform regression rather than classification. We report the mean absolute error (MAE) and the normalized root mean squared error (NRMSE) (defined as $RSME/(k_{max} - k_{min})$) for the full feature set (*All*) and each of the five feature classes defined in Table 2. In addition, we employed two baselines: (1) predict *k* at random, and (2) predict *k* using its marginal distribution. The latter is a stronger baseline since it considers the real distribution of searcher counts in our data. Table 4 (overleaf) reports the performance of the regressor across all experimental runs for all features, the five feature classes, and the two experimental baselines.

We can see from Table 4 that all of the regressors outperform the two baselines (all differences significant with *t*-tests, *p* < 0.001). In addition, the regressor based that uses all features outperforms those regressors based on each class separately, although the performance of some of the individual classes is still fairly strong. Note that the time-of-day features were less effective for this task than they were for the binary classification (pruned temporal model MAE=0.8598, NRMSE=0.1300). Once multiple searchers use the machine, their times may be too intermixed to determine the number of searchers from time alone. Overall, Topical features were found to be most useful in predicting the value of *k* for a machine identifier. To help better understand this result, we explored the feature weights in the predictive models in more detail.

5.4 Feature Weights

One of the advantages of using MART is that we can obtain a list of features learned by the model, ordered by evidential weight. In Table 4 we report the top five most important features from each of the prediction tasks, along with their assigned feature class, and

Table 4. Regression performance for each of the feature classes, ordered by MAE. All differences significant using *t*-tests at *p* < 0.001 for MAE and RMSE for each classifier versus baselines and versus *All*.

Features	MAE	NRMSE
All	0.6377	0.0917
Topical	0.6906	0.1055
Temporal	0.7232	0.1146
Content	0.7490	0.1150
Behavioral	0.8054	0.1204
Referential	0.9784	0.1325
Marginal	1.4799	0.2150
Random	3.8078	0.4652

Table 5. Top five features by evidential weight for the classification and the regression tasks. Feature weights and correlation coefficients for features vs. labels are also shown. Weights are normalized w.r.t. the highest-weighted feature.

	Feature	Class	Weight	r
Classifier	NumTimeBuckets	Temporal	1.0000	+0.180
	FractionWeekday	Temporal	0.6353	+0.444
	FractionQueries_KidsAndTeens	Topical	0.6031	+0.159
	TimeEntropy	Temporal	0.4306	+0.233
	FractionReferenceOtherPerson	Referential	0.3412	+0.149
Regressor	FractionQueries_KidsAndTeens	Topical	1.0000	+0.271
	PageReadability	Content	0.6108	+0.395
	FractionQueries_ShoppingChildren	Topical	0.5550	+0.209
	FractionReferenceOtherPerson	Referential	0.5496	+0.199
	TopicEntropy	Topical	0.3797	+0.143

their weight relative to the most important features: *NumTimeBuckets* (classification) and *FractionQueries_KidsAndTeens* (regression). In addition, to better understand the directionality of the features, we also report in Pearson product moment correlation (r), and the point-biserial correlation in the case of the classifier, between the feature values and the ground truth labels in our dataset.

As we can see in Table 5, the most influential features span many classes, although for the classification task there appears to be more emphasis on Temporal features (as also evidenced by the strong performance of the time-of-day features, shown in Figure 9). However, for the regression task the Topical features are dominant, especially those suggesting the presence of others in the household, in particular children. Indeed, from analyzing the metadata associated with the comScore logs used in our data, we see that the presence of a child in the household is often associated with multiple individuals using the machine. Overall, on 82.1% of machines where a child is in the household we observe multiple searchers, and the phi correlation (r_φ) between child present (1/0) and multi-searcher usage of a machine (1/0) is 0.47 ($p < 0.001$). It seems reasonable that if there is a family linked with the machine then shared usage is more likely. Devising more features associated with family activities (e.g., family vacations, homework searching) may yield even better prediction performance. Topic information may also capture the diversity of interests at more granularity than possible with temporal bucketing (even simply because topics are more numerous), enabling more accurate estimates of searcher counts.

5.5 Summary
We have shown in this section that we can accurately estimate whether a machine identifier comprises multiple searchers and estimate the number of searchers. We convert the output of the regressor to an integral value k' using standard rounding, serving as an estimate of k, the true number of searchers on machine. As stated earlier, being able to estimate the number of searchers from their search behavior is necessary but not sufficient for other tasks such

as assigning incoming search activity to individuals. For that task we must build a *representation* of the search activity for each searcher. To do this, we use a clustering method, where the number of clusters is guided by the output (k') from the regressor in this section. We now describe the application of this estimate to cluster search activity from machine identifiers and attribute new search activity associated with the machine identifier to the correct person.

6. ATTRIBUTING ACTIVITY TO USERS
Given the prediction of multiple searchers on a machine, we tackle following two tasks of attributing observed search activity to individuals: (a) clustering historic search activity guided by the number of estimated searchers k' from the prediction task, (b) automatically assigning new search activity to the most likely individual from historic logs. We now describe each task and model performance.

6.1 Experimental Setting
We used search sessions of individual searchers (defined as in Section 4) for clustering and user history segregation. We can also do this at the level of single queries, but sessions enable the design of richer feature sets. We assume that in practice, the predictors would be chained so that the methods described in this section would only be applied to machine identifiers where the presence of multiple searchers was predicted using the binary classifier from Section 5. For k=1 and k'=1 our performance metrics always equal the baseline (covered later), independent of any attribution technique. As such, in our study we focus on multi-searcher machines ($k > 1$) where our classifier also predicts multi-searcher activity ($k' > 1$).

For measuring the performance of our methods, we use a default baseline which assumes a unique mapping of machine identifier to individual, as that accurately reflects the state of the art methodology for leveraging machine identifiers. When k' is incorrectly predicted to be one for a multi-searcher machine, our attribution techniques would simply fall back to this baseline. Since these attribution tasks require a computationally-expensive clustering of the historic search activity on each machine (and is one of the main motivations for developing the classifier in Section 5), we used a randomly-sampled 5% of machine identifiers (around 90k identifiers) for testing purposes. Further, we split the data for each machine in the test set into historic logs (which corresponds to first 90% of sessions from a machine identifier) and newly arriving queries (corresponding to the latest 10% of the activity). Since our approach is session-based, but we wanted to explore effectiveness for query-based assignment we performed assignment for the first query in the session. Operating at the query level closely resembles how the task of assignment would likely be done in practice in applications such as personalization. Inspecting the person identifier suggests that 97% of search sessions were performed by a single person (the remaining 3% were labeling noise associated with the 30-minute timeout for session demarcation). Accurate assignment for the first query enables personalization for the entire session thereafter.

A key components for the attribution task is measuring "similarity" between two search activities – this measure is critical for both clustering and assignment. Next, we describe the setup for learning this function between two search activities and the set of features used.

6.2 Similarity of Two Search Activities
We begin by representing each of the search activities to be compared as a set of features, informed by the characterization and prediction results in previous sections. We also build a vector representation of the issued query terms, clicked Web domains, search of the activity. Given two search activities denoted by their representative features, we designed around 20 features which capture

Table 6. Features for pairwise similarity of search activity.

Feature	Feature Description
Temporal class	
Diff_Weeks	Diff. in weeks
Diff_DayOfWeek	Diff. in day of week
Sim_Weekday	Both on a weekday
Sim_Weekend	Both on a weekend
Diff_TimeOfDay	Diff. in time of day
Diff_TimeBucket	Diff. in time bucket of a day
Topical class	
Sim_Topic	Jaccard coeff. of ODP categories (clicks)
Sim_QueryCategory	Jaccard coeff. of query categories
Sim_AdultQueryTerms	Jaccard coeff. of adult query terms
Sim_Adult	Both have queries on Adult query category
Sim_Cooking	Both have queries on ODP "Cooking"
Sim_Family	Both have queries on ODP "Family"
Sim_KidsAndTeens	Both have queries on ODP "Kids & Teens"
Sim_Nightlife	Both have queries on ODP "Nightlife"
Sim_ShoppingChild	Both have queries on ODP "Shopping/Children"
Sim_VideoGames	Both have queries on ODP "Video Games"
Behavioral class	
Diff_TotalDuration	Diff. in length of search activity (for session)
Diff_NumQueries	Diff. in # queries
Diff_QueryLength	Diff. in avg. query length (in characters)
Diff_NumClicks	Diff. in total number of clicks
Diff_HistQueryPop	Diff. in avg. historical query popularity
Content class	
Sim_Domains	Jaccard coeff. of web domains (from clicks)
Sim_QueryTerms	Jaccard coeff. of query terms
Sim_Engines	Jaccard coeff. of engine domains (queries)
Diff_QReadingLevel	Diff. in avg. reading level of queries
Referential class	
Sim_RefQueryTerms	Jaccard coeff. of referential query terms
Sim_RefQueries	Both have referential queries

engine domains, ODP and query categories to serve as fingerprint pairwise similarity and differences between two activities for different feature classes, as presented in Table 6. For vector-based features, e.g., a vector of query terms, we used the Jaccard coefficient of two vectors to compute a similarity score, which captures overlapping content in the two vectors. Features such as "*Sim_Weekday*" or "*Sim_Nightlife*" represent a binary value denoting whether both the search activities share the same attribute.

We used MART regression for learning a pairwise similarity function. Since the base prior of two activities belonging to same searcher depends heavily on number of searchers associated with a machine identifier (e.g., similarity score is always 1 for single-searcher machines), we learned different regression models for each k. To create the training data, we used a random sample of 5% of machine identifiers (disjoint from the set used for testing the performance of our attribution tasks) with multiple searchers and split them into five groups containing 2, 3, 4, 5, and 6-10 searchers (the last group comprised a range to provide sufficient data). Next, we learned a regression model for each group separately as follows. We considered every pair of sessions for machines in a group, computed the pairwise features between them and labeled it as 1 if they belong to same searcher, otherwise 0. This labeled data is then used for training the regression model, which learns to predict a real valued number then used as the similarity score between two sessions.

Table 7 shows the most important features ordered by evidential weight. As k increases, we observed that content-based features, primarily based on similarity between query terms and clicked domains, become more prominent. This is likely to happen as multiple searchers may intertwine their search activity more often, adding noise to the temporal signals. Given these regression models, one

Table 7. Top features by evidential weight for pairwise similarity of search activity (machines w/ k=2 and k=5 searchers).

k	Feature	Class	Weight
2	Diff_Weeks	Temporal	1.0000
	Diff_HistoricQueryPopularity	Behavioral	0.3684
	Diff_TotalDuration	Behavioral	0.2997
	Diff_QueryLength	Behavioral	0.2739
	Sim_Engines	Content	0.2687
	Diff_TimeOfDay	Temporal	0.2182
	Diff_NumQueries	Behavioral	0.2030
	Sim_QueryCategory	Topical	0.2009
	Sim_QueryTerms	Content	0.1937
	Diff_NumClicks	Behavioral	0.1936
5	Diff_Weeks	Temporal	1.0000
	Diff_HistoricQueryPopularity	Behavioral	0.4897
	Diff_TotalDuration	Behavioral	0.3030
	Diff_QueryLength	Behavioral	0.2903
	Sim_QueryTerms	Content	0.2738
	Diff_NumClicks	Behavioral	0.2395
	Sim_Engines	Content	0.2366
	Diff_NumQueries	Behavioral	0.2220
	Sim_QueryCategory	Topical	0.2126
	Sim_Domains	Content	0.1827

for each different searcher count (k), we next describe how to use them for our clustering and attribution tasks.

6.3 Clustering Searchers

We now present the results of task of clustering searchers' activities given a history of logs from a machine identifier. To do this, we used the test set of machine identifiers as described in the experimental setup above. We use the output of the regressor from the prediction task in Section 5 to estimate k' for each of the machines. The k' guides both the number of clusters to be used as well as the choice of regression model to use in the similarity computation. For a given machine and predicted k', we first compute pairwise similarity for each pair of sessions and then use k-means clustering with predicted k' as the number of clusters and applying the computed similarity scores as distance metric [46]. Given that we know which searcher is responsible for each observed session, we can use entropy and purity to measure clustering performance [46]. We computed average *entropy* of the clustering solution by computing entropy of each cluster separately based on the distribution of different searchers in that cluster given truth. Our baseline, as discussed earlier, is the default attribution of all the search activity to a single searcher, hence equivalent of having a single cluster for the complete historic logs. We additionally compute the *purity* of the clusters, denoting the fraction of the most representative searcher in a given cluster. A more performant clustering method would yield lower entropy (ideal=0) and higher purity (ideal=1).

Table 8 (overleaf) reports the performance of the clustering task overall and then broken out by the number of searcher on the machine. Our clustering shows significant improvements (using t-tests at $p < 0.001$) on both entropy and purity. Also, the relative improvement increases in magnitude with the true number of searchers suggesting a higher relative benefit from our techniques in segregating searchers. Next, we tackle the task of assignment, i.e., attributing new search activity to one of the searchers who contributed to the search history associated with a machine identifier.

6.4 Assignment of Search Activity

The main challenge assignment is that historic logs are unlabeled, providing no prior information of which activity belong to which searcher. Furthermore, the information about actual number of searchers which contributed to the historic logs is unknown. We tackle this task in following steps: (1) perform clustering on the

Table 8. Average entropy and purity of the clusters obtained by our method vs. baseline. All differences (i.e., drop in entropy, rise in purity) are significant with *t*-tests at *p* < 0.001.

k	Avg. Cluster Entropy	Avg. Cluster Purity	Baseline Entropy	Baseline Purity
All (2–10)	0.552 (–44%)	0.786 (+22%)	0.993	0.644
2	0.551 (–36%)	0.814 (+20%)	0.860	0.676
3	0.542 (–60%)	0.712 (+29%)	1.367	0.553
4	0.601 (–65%)	0.617 (+32%)	1.742	0.467
5	0.635 (–69%)	0.553 (+33%)	2.030	0.415
6–10	0.631 (–72%)	0.515 (+35%)	2.270	0.382

Table 9. Average accuracy and purity of assignment. Baseline assumes unique mapping of machine identifier to searcher. All differences significant with *t*-tests at *p* < 0.001.

k	Accuracy	Purity of Assignment	Baseline
All (2–10)	0.742 (+56%)	0.659 (+39%)	0.475
2	0.771 (+51%)	0.700 (+37%)	0.512
3	0.649 (+89%)	0.512 (+50%)	0.343
4	0.531 (+102%)	0.395 (+50%)	0.263
5	0.451 (+116%)	0.333 (+60%)	0.208
6–10	0.361 (+96%)	0.289 (+57%)	0.184

historic logs as described in previous step to segregate the activity of searchers, and (2) assign the newly arrived query or search activity to one of the clusters. To guide the process of *assigning* a cluster to a searcher, we use the same regression-based similarity function used for clustering (Section 6.3). Based on this function, we first find the activity in the historic logs which is most similar to the new activity. We then assign the new search activity to either (a) the true searcher to which this most similar activity belongs or (b) the cluster to which this most similar activity belongs. These provide different ways to capture error in our assignment.

Given that we have the truth labels, we can measure performance as follows. For (a), we can measure the *accuracy* of the assignment based on how often we assign to the correct individual. For (b), we can measure the *purity of the assignment*. Purity in this context is defined as the proportion of the assigned cluster that comprises the true individual. The overall performance of the assignment is affected by the quality of clustering solution and precision of matching it to the best cluster. An ideal methodology will first perfectly segregate the historic logs into different searchers (cluster purity = 1) and assign the new activity corresponding to the correct cluster (precision of closest assigned searcher = 1). For both measures, we use the same baseline as before, which ignores multi-searcher segregation and assigns the new activity to all historic logs. We note that this baseline deviates from simply $1/k'$ because of imbalance in the search activities of different searchers on a machine.

Table 9 reports the performance of the assignment task overall and then broken down by each of the five groups. We can assign search activity to the correct searcher for over 75% of the newly-arriving queries used in testing. The results show significant gains (*t*-tests at *p* < 0.001) in both measures, compared to the baseline.

6.5 Summary

We have demonstrated that our methods significantly improve the accuracy of search activity attribution, guided by the regressor from Section 5. These results can be used to segregate the activity of different searchers in historic logs. This segregation can help us accurately assign a searcher to new activity in online settings.

7. DISCUSSION AND IMPLICATIONS

We have shown in this paper that searching on shared machines happens frequently, generating signals that could potentially be sub-optimal for applications such as search personalization. We devised methods to accurately label machine identifiers as comprising multiple searchers with good accuracy (even with simply using time-of-day as a feature) and also accurately estimating the *number of searchers* on a machine. We also showed that we could develop methods to attribute search activity to the correct searcher, with accuracy exceeding 70% (for session-level predictions). These findings are promising, but more work is needed to understand model utility in contexts such as personalization and advertising.

While we believe that these applications would benefit from a cleaner historical signal derived from what is likely to be a single

searcher, we need to demonstrate that utility. Beyond personalizing online services, there are also other applications such as long-term search satisfaction modeling [22], protecting privacy between multiple searchers on a single machine, and enhancing search logs with an estimated person identifier to enable more accurate data mining of metrics such as the number of sessions per searcher. However, even without demonstrating these applications directly we make several significant contributions by being the first to identify and characterize the search activity attribution challenge, as well as developing a series of methods to address it effectively.

The comScore data used in this study relied on self-identification. Further work is required to understand the extent to which there are errors in this reporting, its impact on model accuracy, and whether there are ways to address that concern, e.g., by mining patterns from the temporal sequences of searches from machine identifiers and allowing for some degree of noise in the person boundaries.

As people associate more closely with a particular device, we may observe a reduction in shared machine use. Given our findings on the current high prevalence of shared-machine search (over 50% of machines), our work is still valuable. In addition, searchers often sign in to search engines, providing evidence beyond machine identifiers. More research is needed to study the effect of these factors, and learn more about search on shared machines generally.

Overall, we showed that our methods performed well. Even featurizing time-of-day alone was effective in identifying and quantifying shared-machine searching. We also observed strong performance in clustering and attribution, with strong gains over baselines in both tasks. Although we can employ alternative approaches to improve performance in search activity attribution, cost-benefit analyses is required to understand whether any additional complexity is justified given the impressive performance of our methods.

8. CONCLUSIONS

Accurate attribution of online activity is important for providers of online services who seek to personalize their services. We showed that many machine identifiers may reflect the activity of multiple Web searchers, creating a sub-optimal behavioral stream for applications relying on a direct mapping between observed search activity and a single searcher. We showed that there are clues in the long-term behavior that multiple searchers are responsible for the search activity associated to a single machine identifier. Building on this finding, we show that we can accurately predict which machine identifiers have multiple people and estimate with reasonable accuracy the exact number of searchers. This informs the application of clustering methods to predict the total number of searchers for a machine and handle assignment of new activity. Our findings clearly show that we can assign new activity to the correct searchers with good accuracy. Future work involves improving that accuracy, applying the methods to tasks like personalization and recommendation, and experimenting with alternative methods such as component analysis for search activity attribution.

REFERENCES

[1] Amari, S.I., Cichocki, A., and Yang, H.H. (1996). A new learning algorithm for blind signal separation. *Proc. NIPS*, 757–763.

[2] Allen, B. (2000). Individual differences and conundrums of user-centered design. *JASIS*, 51(6): 508–520.

[3] Anand, K., Mathew, G., and Reddy, V. (1995). Blind separation of multiple co-channel BPSK signals arriving at an antenna array. *IEEE Signal Processing Letters*, 2: 176–178.

[4] Barabasi, A.L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039): 207–211.

[5] Bennett, P.N., Svore, K., and Dumais, S.T. (2010). Classification-enhanced ranking. *Proc. WWW*, 111–120.

[6] Bhavnani, S. (2001). Important cognitive components of domain-specific search knowledge. *Proc. TREC*, 571–578.

[7] Bilenko, M. and Richardson, M. (2011). Predictive client-side profiles for personalized advertising. *Proc. SIGKDD*, 413–421.

[8] Buscher, G., White, R.W., Dumais, S.T., and Huang, J. (2012). Large-scale analysis of individual and task differences on search result page examination strategies. *Proc. WSDM*, 373–424.

[9] Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2003). Visualization of navigation patterns on a web site using model based clustering. *Data Mining and Knowledge Discovery*, 7: 399–424.

[10] Cardoso, J.F. (1998). Blind signal separation: statistical principles. *Proc. IEEE*, 86(10): 2009–2025.

[11] Chen, Y., Pavlov, D., and Canny, J. (2009). Large-scale behavioral targeting. *Proc. SIGKDD*, 209–218.

[12] Collins-Thompson, K., and Callan, J. (2004). A language modeling approach to predicting reading difficulty. *Proc. HLT*, 193–200.

[13] Comon, P. (1994). Independent component analysis: a new concept? *Signal Processing*, 36(3): 287–314.

[14] Dasgupta, A., Gurevich, M., Zhang, L., Tseng, B., and Thomas, A.O. (2012). Overcoming browser cookie churn with clustering. *Proc. WSDM*, 83–92.

[15] Dou, Z., Song, R., and Wen, J.R. (2007). A large-scale evaluation and analysis of personalized search strategies. *Proc. WWW*, 581–590.

[16] Downey, D., Dumais, S.T., and Horvitz, E. (2007). Models of searching and browsing: languages, studies, and applications. *Proc. IJCAI*, 2740–2747

[17] Dumais, S., Buscher, G., and Cutrell, E. (2010). Individual differences in gaze patterns for Web search. *Proc. IIiX*, 185–194.

[18] Dupret, G. and Piwowarski, B. (2008). A user browsing model to predict search engine click data from past observations. *Proc. SIGIR*, 331–338.

[19] Fawcett, T. and Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3): 291–316.

[20] File, T. (2013) *Computer and Internet Use in the United States*. http://www.census.gov/prod/2013pubs/p20-569.pdf

[21] Fulgoni, G.M. (2005). *The "Professional Respondent" Problem in Online Survey Panels Today*. Slides online at: http://www.sigmavalidation.com/tips/05_06_02_Online_Survey_Panels.ppt (Downloaded on October 3, 2013).

[22] Hu, V, Stone, M., Pedersen, J., and White, R.W. (2011). Effects of search success on search engine re-use. *Proc. CIKM*, 1841–1846.

[23] Friedman, J.H., Hastie, T., and Tibshirani, R. (1998). Additive logistic regression: A statistical view of boosting. Technical Report, Department of Statistics, Stanford University.

[24] Joachims, T. (2002). Optimizing search engines using click-through data. *Proc. SIGKDD*, 133–142.

[25] Kobsa, A. (2007). Privacy-enhanced personalization. *CACM*, 50(8): 24–33.

[26] Krause, A. and Horvitz, E. (2008). A utility-theoretic approach to privacy and personalization. *Proc. AAAI*, 1181–1188.

[27] Lau, T. and Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement. *Proc. UM*, 119–128.

[28] MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Symposium on Math, Statistics, and Probability*, 281–297.

[29] Matthijs, N. and Radlinski, F. (2011). Personalizing web search using long term browsing history. *Proc. WSDM*, 25–34.

[30] Richardson, M. (2009). Learning about the world from long-term query logs. *ACM TWEB*, 2(4): 21.

[31] Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. *Proc. ASIS*, 82–86.

[32] Shen, X., Dumais, S., and Horvitz, E. (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102–1103.

[33] Shen, X., Tan, B., and Zhai, C.X. (2005). Implicit user modeling for personalized search. *Proc. CIKM*, 824–831

[34] Sontag, D., Collins-Thompson, K., Bennett, P.N., White, R.W., Dumais, S., and Billerbeck, B. (2012). Probabilistic models for personalizing web search. *Proc. WSDM*, 433–442.

[35] Tan, B., Shen, X., and Zhai, C. (2006). Mining long-term search history to improve search accuracy. *Proc. SIGKDD*, 718–723.

[36] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proc. SIGIR*, 449–456.

[37] Teevan, J., Liebling, D.J., and Geetha, G.R. (2011). Understanding and predicting personal navigation. *Proc. WSDM*, 85–94.

[38] Thatcher, A. (2008). Web search strategies: The influence of web experience and task type. *IP&M*, 44(3): 1308–1329.

[39] White, R.W. and Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. *Proc. SIGIR*, 255–262.

[40] White, R.W. and Drucker, S. (2007). Investigating behavioral variability in web search. *Proc. WWW*, 21–30.

[41] White, R.W., Bailey, P., and Chen, L. (2009). Predicting user interests from contextual information. *Proc. SIGIR*, 363–370.

[42] White, R.W., Dumais, S.T., and Teevan, J. (2009). Characterizing the influence of domains expertise on web search behavior. *Proc. WSDM*, 132–141.

[43] White, R.W., Bennett, P.N., and Dumais, S.T. (2010). Predicting short-term interests using activity-based search context. *Proc. CIKM*, 1009–1018.

[44] Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., and Li, H. (2010). Context-aware ranking in web search. *Proc. SIGIR*. 451–458.

[45] Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., and Chen, Z. (2009). How much can behavioral targeting help online advertising? *Proc. WWW*, 261–270.

[46] Zhao Y. and Karypis, G. (2002). Criterion functions for document clustering: Experiments and analysis. *Proc. CIKM*, 515–524.