ABSTRACT

Alice’ submitting one web search per five minutes, for three hours in a row—is it normal? How to detect abnormal search behaviors, among Alice and other users? Is there any distinct pattern in Alice’s (or other users’) search behavior? We studied what is probably the largest, publicly available, query log, containing more than 30 million queries from 0.6 million users. In this paper, we present a novel, user-and group-level framework, M3A: Model, MetaModel and Anomaly detection. For each user, we discover and explain a surprising, bi-modal pattern of the inter-arrival time (IAT) of landed queries (queries with user click-through). Specifically, the model Camel-Log is proposed to describe such an IAT distribution; we then notice the correlations among its parameters at the group level. Thus, we further propose the metamodel Meta-Click, to capture and explain the two-dimensional, heavy-tail distribution of the parameters. Combining Camel-Log and Meta-Click, the proposed M3A has the following strong points: (1) the accurate modeling of marginal IAT distribution, (2) quantitative interpretations, and (3) anomaly detection.

1. INTRODUCTION

‘Alice’ is submitting one web search per five minutes, for three hours in a row—is it normal? How to detect abnormal search behaviors, among Alice and other users? Is there any distinct pattern in Alice’s (or other users’) search behavior? These three questions serve as the motivations of this work.

Conventionally, each of Alice’s queries is assumed (1) to be submitted independently and (2) to follow a constant rate \( \lambda \), which results in a simple and elegant model, Poisson process (PP). PP generates independent and identically distributed (i.i.d.) inter-arrival time (IAT) that follows an (negative) exponential distribution [8]. In reality, however, does PP accurately model her search behavior?

To answer this question, we investigate a large, industrial query log that contains more than 30 million queries submitted by 0.6 million users. Figure 1 illustrates the histogram of a user’s IAT. The temporal resolution is one second. As Figure 1(a) shows, this distribution has a “heavy tail” as opposed to an (negative) exponential distribution whose tail decays exponentially fast. In the logarithmic scale as Figure 1(b) shows, surprisingly, two distinct modes (denoted as \( M_1 \) and \( M_2 \)) with approximately symmetric shapes can be seen. This distribution (or a mixture of distributions) clearly does not follow an (negative) exponential distribution, which has a strictly right-skewed shape in logarithmic scale and therefore cannot depict such shapes. This phenomenon suggests that the assumptions of PP rarely hold, since the arrival rate may change, or certain queries may be submitted depending on the previous queries.

In this paper we aim at solving the following problems:

- **P1:** Pattern discovery and interpretation. Is there any pattern in the IAT on Alice’s behalf?
- **P2:** Behavioral modeling. How to characterize the marginal distribution of IAT?
- **P3:** Anomaly detection. Given IAT from ‘Bob,’ how to determine whether his behavior is abnormal from Alice and other users?

The answers to the above questions are exactly the contributions brought by the proposed M3A:

- **A1:** Pattern discovery and interpretation. One key observation of IAT is provided: a bi-modal \((M_1, M_2)\) distribution with \( M_1 \) referred as in-session whereas \( M_2 \) is referred as take-off (e.g., sleep time) query.
- **A2:** Behavioral modeling. Specifically, we propose:
  - “Camel-Log” to parametrically characterize Alice’s (or any person’s) IAT by mixing two heavy-tail distributions.
  - “Meta-Click” to describe the joint probability of two parameters of Camel-Log by using a lesser-known tool of Copula.
- **A3:** Anomaly detection. Camel-Log generates IAT with the same statistical properties as in the real data shown in Figure 1(b), and Meta-Click can detect abnormal users as in Figure 1(c)(d).

The remainder of this paper is organized as follows. Section 2 provides the problem definition. Section 3 details the user-level model Camel-Log and Section 4 details the group-level metamodel Meta-Click. Section 5 provides the usage of M3A. Section 6 surveys the previous work. Finally, Section 7 concludes this paper.

2. PROBLEM DEFINITION

In this work, we use a large-scale, industrial query log released by AOL [14], which is essentially a Google query log since AOL.

---

*Dr. Da-Cheng Juan is now with Google Inc.
searches are powered by Google. The basic statistics of this query log are provided here:

- Duration: three months, from March 1st to May 31st, 2006.
- 36 millions queries submitted from 657,000 users:
  - 19 millions queries WITH click-through (referred as landed queries).
  - 17 millions queries WITHOUT click-through (referred as orphan queries).
- The temporal resolution is 1 second.

2.1 Terminology and problem formulation

Table 1 provides the symbols and the corresponding definitions used throughout this paper. By the convention in statistics, random variables are represented in upper-case (e.g., $M$) and the corresponding values (e.g., $m$) are in lower-case.

As mentioned in Section 1, we aim at solving the following three problems:

**Problem 1 (Pattern discovery and interpretation).** Given each user ID and the time stamp of each query, find and interpret the most distinct pattern sufficient to characterize the IAT distribution of each user.

**Problem 2 (Behavioral modeling).** Given the pattern found in P1, design:

1. A model (and a metamodel) that matches the statistical properties of the empirical data.
2. The parameters (and the hyper-parameters).

**Problem 3 (Anomaly detection).** Given:

1. The model (and metamodel) from P2.
2. The time stamp of each query from a user.

Determine if her/his query behavior in terms of IAT is abnormal.

2.2 Observation on non-landed queries: “orphan queries”

In Figure 2 notice that certain users (marked by the red rectangle) have submitted more than 1,000 queries but clicked through very few (less than 100, or even zero!) of them, resulting in abnormally many of orphan queries. Another obvious evidence is: these orphan queries usually submitted (a) consecutively and (b) with the same keyword, leading to a clear robotic behavior. Therefore, we provide the following qualitative observation.

**Observation 1 (Orphan queries).** Users who have submitted many (usually more than 1,000) queries but clicked through very few (less than 100) of them are abnormal.

Furthermore, one user (circled by red) in the upper-right corner of Figure 2 has submitted more queries (by two order of magnitudes, $\approx 130,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$) than typical users ($\approx 30,000$), with the longest IAT of only 20 minutes (no sleep time). Clearly, this user is suspicious and therefore an anomaly.

After being able to detect obvious anomalies with orphan queries, we again ask the major motivating question (as mentioned in Section 1): “How frequently does ‘Alice’ submit a web query and click through the search results?” Starting immediately, we ignore orphan queries and focus on the IAT of landed queries.

Figure 1: Patterns and anomalies with M3A: (a) Histogram of inter-arrival time (IAT) for a single user in linear scale. No prevailing patterns are shown. (b) Logarithmic binning (equally-spaced in log-scale) of IAT with Camel-Log fit. A bi-modal distribution can be seen: $M_1$ at 5 minutes (typical inter-query time) and $M_2$ at hours (typical time between sessions). (c) illustrates group-level analysis with scatter plot of the ratio (in-session/take-off queries) vs. the median of in-session intervals. Anomalies are spotted: anomalies (circled by red) cannot be detected by using only the marginal PDF of X-variable, whereas anomalies (marked by the red rectangle) cannot be detected by using the Y-variable. (d) shows an automated way of spotting anomalies through Meta-Click: the blue deviants (within red circles/boxes) correspond to the outliers (in circles/boxes) in (c).

Figure 2: Orphan queries. Queries without following through are suspicious (see the red box). One user (circled by red) has submitted $\approx 130,000$ queries, with the longest IAT of only 20 minutes (no sleep time).
Table 1: Symbols and definitions

| Symbol     | Definition                                                                 |
|------------|---------------------------------------------------------------------------|
| IAT        | Inter-arrival time                                                        |
| \(t_{i,j}\) | IAT between \(j^{th}\) and \((j + 1)^{th}\) query submitted by user \(i\).
| \(F_T(\cdot)\) | Cumulative distribution function (CDF) for: (a) the random variable \(T\) or (b) the distribution \(T\) |
| \(f_T(\cdot)\) | Probability density function (PDF) for: (a) the random variable \(T\) or (b) the distribution \(T\) (e.g., \(f_{\mathcal{LL}}\) is the PDF of log-logistic) |
| \(\mathcal{LL}\) | Log-logistic distribution: a skewed (in linear scale), heavy-tail distribution |
| Camel-Log  | Proposed mixture of two log-logistic distribution: modeling marginal IAT    |
| Meta-Click | Proposed 2-d log-logistic distribution using Gumbel’s copula: metamodeling the parameters of Camel-Log |

Symbols used by Camel-Log

- \(\alpha_{in}, \beta_{in}\): Parameters: median and shape of log-logistic distribution (for modeling in-session IAT)
- \(\alpha_{off}, \beta_{off}\): Parameters: median and shape of log-logistic distribution (for modeling take-off IAT)
- \(\theta\): Proportion parameter: \(\theta \in [0,1]\) for in-session IAT, and \((1 - \theta)\) for take-off IAT

Symbols used by Meta-Click

- \(R\): Random variable representing the ratio of in-session and take-off IAT: \(R \triangleq \theta/(1 - \theta)\)
- \(M\): Random variable representing the log-median of in-session IAT: \(M \triangleq \log(\alpha_{in})\)
- \(\alpha_R, \beta_R\): Hyper-parameters: median and shape of log-logistic distribution (for modeling \(R\))
- \(\alpha_M, \beta_M\): Hyper-parameters: median and shape of log-logistic distribution (for modeling \(M\))
- \(C(\cdot, \cdot)\): Copula: Joint CDF of two random variables considering their dependency \([0,1] \times [0,1] \rightarrow [0,1]\)
- \(\eta\): Parameter in Gumbel’s copula that captures correlations between random variables \(R\) and \(M\)

3. SINGLE USER ANALYSIS: Camel-Log

In this section, we first detail the proposed Camel-Log distribution (Section 3.1), provide validations (Section 3.2) and give comparisons with other well-known models (Section 3.3). For convenience, we preview the mathematical form of Camel-Log here:

\[
f_{\text{Camel-Log}}(t) = \theta \cdot f_{\mathcal{LL}}(t; \alpha_{in}, \beta_{in}) + (1 - \theta) \cdot f_{\mathcal{LL}}(t; \alpha_{off}, \beta_{off})
\]

where \(t \geq 0\), \(f_{\mathcal{LL}}(\cdot)\) stands for the probability density function (PDF) of log-logistic (\(\mathcal{LL}\)) distribution as shown in Eq(2).

3.1 Camel-Log distribution

The main idea of Camel-Log is to use a mixture of two log-logistic (\(\mathcal{LL}\)) distributions to model the bi-modal pattern in Figure 1(b). \(\mathcal{LL}\) is a skewed (in linear scale), power-law-like (heavy-tail) distribution, and there are two reasons for the choice of \(\mathcal{LL}\): (a) it outperforms competitors (see Section 3.3); (b) it has an intuitive explanation (the longer a person has waited, the longer (s)he will wait). \(\mathcal{LL}\) has been used successfully for modeling the IAT of the Internet communications of humans, such as posts on web blogs and comments on the Youtube.\(^3\) We remind its definition here:

**Definition 1 (Log-Logistic distribution).** Let \(T\) be a non-negative continuous random variable and \(T \sim \mathcal{LL}(t; \alpha, \beta)\). The CDF of a log-logistically distributed \(T\) is given as:

\[
F_{\mathcal{LL}}(t; \alpha, \beta) = \frac{1}{1 + (t/\alpha)^{\beta}}
\]  

(1)

where \(\alpha > 0\) is the median (or called scale parameter), and \(\beta > 0\) is the shape parameter. The support \(t \in [0, \infty)\). The PDF of \(T\) is given as:

\[
f_{\mathcal{LL}}(t; \alpha, \beta) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{[1 + (t/\alpha)^{\beta}]^2}
\]  

(2)

With the knowledge of \(\mathcal{LL}\), we present the definition of the proposed Camel-Log distribution:

**Definition 2 (Camel-Log distribution).** Let \(T\) be a non-negative random variable following Camel-Log distribution. The probability density function (PDF) can be written as:

\[
f_{\text{Camel-Log}}(t) = \theta \cdot f_{\mathcal{LL}}(t; \alpha_{in}, \beta_{in}) + (1 - \theta) \cdot f_{\mathcal{LL}}(t; \alpha_{off}, \beta_{off})
\]  

(3)

where \(t \geq 0\), \(\theta \in [0,1]\), \(\alpha_{in}, \beta_{in}, \alpha_{off}, \beta_{off} > 0\).

The proposed Camel-Log distribution has the following properties:

- A mixture of two \(\mathcal{LL}\) (heavy-tail) distributions to qualitatively describe: in-session and take-off IAT.
- Five parameters to characterize ‘Alice’s’ search behavior:
  - \(\theta\) controls the proportion of in-session and take-off IAT.
  - \(\alpha_{in}\) represents the median of in-session IAT.
  - \(\beta_{in}\) is the “concentration”\(^3\) of in-session IAT.
  - \(\alpha_{off}\) represents the median of take-off IAT.
  - \(\beta_{off}\) is the concentration of take-off IAT.

Camel-Log distribution seems to model the marginal distribution of IAT very well, at least for ‘Alice’ shown in Figure 1(b), and also provides intuitive interpretations. But we still have the following questions:

\(^3\)The reciprocal of \(\beta_{in}\) represents (approximately) the standard deviation of \(\mathcal{LL}\).
Also from Figure 3, we provide the following observation:

- Is Camel-Log sufficiently general and accurate to model and interpret other people’s search behavior?
- Even so, does $\mathcal{CL}$ outperform other famous distributions, say Exponential or Pareto (power-law)?

The answers to both questions are yes, and the details are provided in the following two sections.

### 3.2 Validation against empirical data

Figure 3 illustrates the empirical IAT from 12 most ‘prolific’ users. Each sub-figure shows the marginal distribution of IATs (in logarithmic binning) from a user, and the red curve is depicted by fitting a Camel-Log distribution via expectation maximization (EM).

Figure 4 provides the Quantile-Quantile plot (Q-Q plot) between the empirical IAT and the samples randomly drawn from the fitted Camel-Log distribution. In each sub-figure, X axis represents the IAT from a user and Y axis are the samples drawn from the fitted Camel-Log distribution. $45^\circ$ line is ideal, meaning that the empirical data and the fitted samples follow the same distribution.

As it can be seen, in each sub-figure the majority of quantiles are matched very well by the proposed Camel-Log distribution.

3.3 Why not other well-known distributions?

in-session IAT is about five minutes, whereas the median of take-off IAT is approximately seven hours.

There are two types of IAT: in-session and take-off. The median of in-session IAT is about five minutes, which approximately represents the duration when a user is interested in the query results. On the other hand, the IAT of take-off queries is longer, ranging from tens of minutes (e.g., lunch break), hours (e.g., sleep time), to days (e.g., weekends). The median of take-off IAT is approximately seven hours, which corresponds to sleep time very well.

More validations are provided by Figure 4. For each user, Figure 4 provides the Quantile-Quantile plot (Q-Q plot) between the empirical IAT and the samples drawn from the fitted Camel-Log distribution. $45^\circ$ line is ideal, meaning that the empirical data and the fitted samples follow the same distribution.

As it can be seen, in each sub-figure the majority of quantiles are matched very well by the proposed Camel-Log distribution.

By now we have strong evidences supporting the goodness of fit for Camel-Log; we still need to answer the question: why not using a mixture of other well-known “named” distributions, say Exponential or Pareto (power-law)?
Exponential mix.

Pareto mix.

66%

78%

Exponential mix.

Pareto mix.

on 78% of the users (compared to Exponential mixture), and more
candidates. The proposed Camel-Log achieves a higher log-likelihood
Log explains better (achieves higher likelihood), compared to other
Pareto mixture does not fit at all (with constantly low p-values).
true model; exponential mixture fits well but not as close, whereas
Figure 5, the proposed Camel-Log is the candidate closest to the
form(0,1) distribution, depicted by the 45◦
date distribution. If

H0: the user’s IAT follows the fitted can-
dition distribution. If H0 is true, the p-value will follow a uni-
form(0,1) distribution, depicted by the 45◦
line.

Figure 5 also shows that Camel-Log fits the Reddit dataset well
by Q-Q plot. Notice that the majority of quantiles match very well.
Therefore, the generality of the proposed Camel-Log is demon-
strated: Camel-Log fits and explains multiple datasets (both Google
queries and Reddit posts).

Since Camel-Log characterizes each user’s search behavior by
five parameters, we ask: how to use these parameters, specifically
the ratio (R) and the log-median (M), to detect anomalies as Figure
(c)?

4. GROUP-LEVEL ANALYSIS: Meta-Click

Are there regularities, in the parameters of all the users? It turns
out that yes, some of the parameters are correlated. The two that
show a stronger correlation are the ratio R (≜ \frac{\alpha}{\theta}) and the
log-median M (≜ \log(\alpha_J)). Thus, our goal is to model the joint
distribution.

Jumping ahead, given that both their marginals follow \( LL \) (see
Section 5.1), how should we combine them, to reach a joint distri-
bution that models Figure 1(c)? The main idea is to use a powerful
statistical tool, Copulas (see Section 4.3). For convenience, the

1Compared to Akaike information criterion (AIC).
2Given any two estimated models, the model with the lower value of
BIC is the one to be preferred.
3http://www.reddit.com/
4The dataset contains 16,927 unique users; for each user, we collect
the timestamp of 500 his/her posts.
Figure 6: Camel-Log fits the Reddit dataset (marginal PDF). Each sub-figure shows the marginal distribution of IATs and the proposed Camel-Log fitting results (in red). Notice that Camel-Log fits well. Further notice the consistency of the bimodal (in-session, take-off) behaviors.

Figure 7: Camel-Log fits the Reddit dataset (Q-Q plot). Each sub-figure shows the Q-Q plot (ideal: 45° line) between the real data and the samples randomly drawn from the fitted Camel-Log. Notice that the majority of quantiles match very well.

With the parameters extracted by Camel-Log (specifically, $\theta$ and $\alpha_{cin}$ for each user), we define two random variables that are particularly useful for anomaly detection:

- **Ratio**: $R \triangleq \theta/(1-\theta)$ that represents approximately how many “query and click”s happening within a search session (in-session) v.s. take-off.
- **Log-median**: $M \triangleq \log(\alpha_{cin})$ represents the median of in-session IAT in log scale.

Intuitively, $R$ and $M$ represent an aggregate behavior, in terms of a statistical distribution of parameters (specifically, $\theta$ and $\alpha_{cin}$) used to characterize each user. Figure 6 illustrates the marginal distribution of $R$ in (a) and $M$ in (d), respectively. Note that all the LLC fittings are done by using Maximum Likelihood Estimate (MLE).

To better examine the distribution behavior both in the head and tail, we propose to use the Odds Ratio (OR) function.

**Lemma 1 (Odds Ratio).** In logarithmic scale, $OR(t)$ has a linear behavior, with a slope $\beta$ and an intercept $(-\beta \log \alpha)$, if $T$ follows Log-logistic distribution. From the definition of OR function, we have:

$$OR(t) = \frac{F_R(t)}{1 - F_R(t)} = \left(\frac{t}{\alpha}\right)^\beta$$

(4)

$$\Rightarrow \log OR(t) = \beta \log(t) - \beta \log \alpha$$

Figure 6(c)(f) show the OR of $R$ and $M$, respectively. For both random variables, their ORs seem to entirely follow the linear line, which serves as another evidence that their marginal distributions follow LLC. K-S tests are also conducted for both $R$ and $M$; under 95% confidence level, we retain the null hypothesis: $R$ (and $M$) follows the fitted LLC.

**Observation 3 (Common User Behavior).** The mode of the ratio $R$ is approximately three, which suggests a common user behavior: “click-click-click—taken off—then click (new session)”.

The marginals of $R$ and $M$ follow LLC, but how about their two-dimensional joint distribution ($F_{R,M}$)? Can we use a multivariate normal (MVN) distribution to describe them?

**4.2 Why not multivariate normal (MVN)?**

Modeling multivariate distribution is a rather challenging task. One popular method is to use a multivariate normal (MVN) distri-
Figure 8: Marginal distributions follow \( \mathcal{L}\mathcal{L} \) distributions: (a) Marginal distribution of \( R \) and the \( \mathcal{L}\mathcal{L} \) fitting. (b) Q-Q plot between empirical \( R \) and fitted \( \mathcal{L}\mathcal{L} \). (c) Odds Ratio (OR) between empirical \( R \) and fitted \( \mathcal{L}\mathcal{L} \). (d)(e)(f) provide the corresponding plots for \( M \). In (c), the OR of \( R \) seems to entirely follow the linear line, which serves as another evidence that its marginal distribution follows a \( \mathcal{L}\mathcal{L} \). The same statement also holds for (d). K-S tests are conducted for both \( R \) and \( M \); under the 95% confidence level, we retain the null hypothesis: the empirical data follows the fitted \( \mathcal{L}\mathcal{L} \).

### 4.3 A crash introduction to Copulas

In statistics, Copulas are widely-used to model a multivariate, joint distribution considering the dependency structures between random variables (e.g., \( R \) and \( M \)). The main concept of Copulas is to associate univariate marginals (e.g., \( F_R, F_M \)) with their full multivariate distribution. Here, we remind the mathematical definition of copula as below:

**Definition 3 (Copula).** A copula \( C(u, v) \) is a dependence function defined as:

\[
C : [0, 1] \times [0, 1] \rightarrow [0, 1]
\]  \hspace{1cm} (5)

Given two random variables \( R, M \) and their marginal CDFs \( F_R, F_M \), a copula \( C(u, v) \) generates a joint CDF that captures the correlation between \( R \) and \( M \): \( F_{R,M}(r, m) = C(F_R(r), F_M(m)) \).

In theory, Copulas can capture any type of dependency between variables: positive, negative, or independence. The existence of such Copula is guaranteed by Sklar’s Theorem.\(^7\)

One type of Copulas is very popular in modeling joint distribution of random variables with heavy tails: **Gumbel Copula**. We remind the definition of Gumbel Copula as below:

**Definition 4 (Gumbel Copula).** A Gumbel Copula is defined as:

\[
C(u, v) = e^{-[\phi(u) \eta + \phi(v)]^{1/\eta}}
\]  \hspace{1cm} (6)

where \( \eta \geq 1 \) and \( \phi(.) = -\log(.) \).

\(^7\)The details of Sklar’s theorem can be found in [17].
Figure 9: Meta-Click matches real data. (a)-(c): contour plots for Meta-Click (with various $\eta$). (d): real data. All plots are $R$ v.s. $M$. In (b), $\eta = 1.12$, which is the value estimated from the real data. Notice how well (b) matches (d).

Notice that $C(u, v) = u \cdot v$ when $\eta = 1$, indicating that $u, v$ are independent.

With this tool, we are ready to proceed to the proposed Meta-Click.

### 4.4 Proposed Meta-Click

The goal of Meta-Click is to model the joint distribution of $R$ and $M$. As the results presented in Section 4.1, their marginals follow $LL$. By using Gumbel Copula, we present the definition of the proposed Meta-Click here:

**Definition 5 (Meta-Click).** Let $R$ and $M$ be non-negative random variables following Meta-Click distribution, the CDF of their joint distribution is:

$$F_{\text{Meta-Click}}(r, m; \eta, \alpha_R, \beta_R, \alpha_M, \beta_M) = e^{-\left(\log(1+(r/\alpha_R)^{-\beta_R}))\eta + \log(1+(m/\alpha_M)^{-\beta_M})\eta\right)^\eta}$$

where $r, m \geq 0$, $\eta \geq 1$, $(\alpha_R, \beta_R)$, $(\alpha_M, \beta_M)$ are the hyper-parameters used in $F_{\text{CL}}(r)$ and $F_{\text{CL}}(m)$, respectively.

In this work, $\eta$ in Eq. (7) is estimated by Kendall tau correlation [10]; the values of $(\alpha_R, \beta_R)$, $(\alpha_M, \beta_M)$ are estimated by using MLE as mentioned in Section 4.1.

We now show that the proposed Meta-Click distribution preserves the characteristics in the marginal distributions of each random variable:

**Lemma 2 (Marginals of Meta-Click are LL).** We prove this by taking the limit of $r$ to infinity:

$$\lim_{r \to \infty} F_{\text{Meta-Click}}(r, m) = F_M(m; \alpha_M, \beta_M) = \frac{1}{1 + (m/\alpha_M)^{-\beta_M}}$$

Therefore, $M \sim LL(\alpha_M, \beta_M)$. We can show $R \sim LL(\alpha_R, \beta_R)$ in a similar manner.

Figure 9(a) (b) (c) illustrate three contour plots of the proposed Meta-Click with setting $\eta$ to various values, whereas Figure 9(d) provides the contour plot from the empirical data. The contour plot in (b) seems to match the empirical data qualitatively well.

### 5. M3A: Practitioners’ Guide

We provide the step-by-step guide to apply the proposed M3A for behavioral modeling and anomaly detection:

- **Camel-Log at user level**: given a user’s IAT, use Camel-Log to characterize their marginal IAT distribution with five parameters ($\theta, \alpha_R, \beta_R, \alpha_M, \beta_M$) in Eq(6).
- **Meta-Click at group level**: given each user’s $\theta$ and $\alpha_M$ from the previous step, convert them into ratio $R$, log-median $M$ and then use Meta-Click presented in Eq(7) to estimate Copula parameter $\eta$ for the two-dimensional heavy-tail distribution.
- **Anomaly detection**: given a user’s $R$ and $M$, calculate its likelihood by using Meta-Click.

Figure 10 presents the anomalies detected by M3A. Figure 10(b) provides “rank-weirdness” plot: users are presented in a “least likely first” order, by using the likelihood of observing their $R$ and $M$ calculated by Meta-Click. All users fit on a line, except the first seven users who have tiny likelihoods. As a comparison, the green line shows a synthetic set of users by using Eq(6). Notice that none of the “green” users exhibits such tiny likelihoods; further notice that those seven users indeed correspond to outliers in ($R, M$) space, where we enclose them in a red box and two red ellipses for visual clarity in Figure 10(a).

Figure 10(c) further illustrates an abnormally-active user detected by M3A. Notice the disproportion between in-session and take-off (the ratio $R \approx 30$), which is ten times higher compared to a typical user’s (around 3).

### 6. Related Work

Many prior papers have attempted to model the temporal, Internet-based activities of humans:

- **Internet-based, temporal data.** Vaz de Melo et al. [20] have proposed a self-learning process to generate IAT following $LL$ distributions for modeling the Internet-based communications of humans. Becchetti et al. [3] and Castillo et al. [5] have proposed novel graph-based algorithms for Web spam detection. Meiss et al. [12] have demonstrated that client-server connections and traffic flows exhibit heavy-tailed probability distributions lacking any typical scale. Münz et al. [13] have presented a flow-based anomaly detection scheme based on the K-means clustering. Gupta et al. [9] provides a comprehensive survey on outlier detection for temporal data. Veeco et al. [19] have proposed a time-based collective factorization for monitoring news. Xing et al. [21] have proposed to use local shapelets for early classification on time-series data. Ratanamahatana et al. [15] gives a high-level survey of time-series data mining tasks, with an emphasis on time series representations. Furthermore, point
processes, time series and inter-arrival time analysis have attracted huge interests, with multiple textbooks (Keogh et al. [4]).

- **Human activities.** Shie et al. [18] has proposed a new algorithm (IM-Span) for mining user behavior patterns in mobile commerce environments. Saveski et al. [16] has adapted active learning to model the web services. Barabasi [2] models and explains human dynamics with heavy-tail distributions. Liu et al. [11] have provided a Weibull analysis of Web dwell time, to discover human browsing behaviors. Sarma et al. [6] provides a fine tutorial on personalized search.

Table 3 summarizes the comparison among several popular methods. As Table 3 shows, this is the only work focusing on the surprising pattern of web query IAT: in-session and take-off, and proposing a new framework M3A to (a) match and explain this pattern, and (b) detect anomaly. To the best of our knowledge, this is the first work to use log-logistic distributions and the Copulas (as a metamodel) to describe the IAT of web queries.

### 7. CONCLUSION

In this paper, we answer the motivational questions mentioned in the Introduction: ‘Alice’ is submitting one web search per five minutes, for three hours in a row—is it normal? How to detect abnormal search behaviors, among Alice and other users? Is there any distinct pattern in Alice’s (or other users’) search behavior?

We conclude this paper by bringing the answers to these questions:

- **A1: Pattern discovery and interpretation.** One key observation of IAT is provided: a bi-modal distribution with the interpretation of in-session and take-off behaviors.
- **A2: Behavioral modeling.** Specifically, we propose:
  - “Camel-Log” to parametrically characterize Alice’s (or any person’s) IAT by mixturing two log-logistic distributions.
  - “Meta-Click” to describe the joint probability of two parameters of Camel-Log by using Gumbel Copula.
- **A3: Anomaly detection.** Camel-Log generates IAT with the same statistical properties as in the real data, and Meta-Click can detect abnormal users by examining their search behaviors.

Finally, we provide a practitioners’ guide for M3A, and illustrate its power via “rank-weirdness” plot as in Figure 10(b). M3A exactly pin-points the outliers that a human would spot: the points in red circles/boxes, in Figure 10(a).

### 8. REFERENCES
Appendix

Kolmogorov-Smirnov (K-S) test
Kolmogorov-Smirnov test (K-S test) is a non-parametric statistical test for testing the equality of two probability distributions. The null hypothesis assumes the samples are drawn from the given continuous distribution. Mathematically, the Kolmogorov-Smirnov test statistic is defined as:

\[ D_n = \sup_x |F_n(x) - F(x)| \]

where \( F_n(x) \) is the empirical distribution estimated from the sample population, and \( F(x) \) is the cumulative distribution function (CDF) of the given probability distribution. Under the null hypothesis, \( \sqrt{n}D_n \) converges to the Kolmogorov distribution. Hence, the risk region of Kolmogorov-Smirnov test is \( \sqrt{n}D_n > K_n \), where \( K_n \) satisfies that \( P(K > K_n) = 1 - \alpha \), \( K \) follows Kolmogorov distribution.

Bayesian information criterion (BIC)
Bayesian information criterion (BIC) is a criterion for model selection. In model selection, the criterion purely based on log-likelihood is likely leading to over-fitting. BIC is a penalized version of log-likelihood. Mathematically,

\[ BIC = -2L + k \ln(n) \]

where \( L \) is log-likelihood, \( k \) is the number of parameters, and \( n \) is number of observations. Hence, minimizing BIC tends to select model with less parameters (parsimony).

Kendall tau in Gumbel copula
Kendall tau correlation \( \eta \) measures the dependency between two random variables. Given random variables \( X \), \( Y \) and \( n \) pairs of their observations, \((x_1, y_1), \ldots, (x_n, y_n)\), a pair of observations \((x_i, y_i)\) and \((x_j, y_j)\) is called concordant if \((x_i - x_j)(y_i - y_j) > 0\). Likewise, the pair is called discordant if \((x_i - x_j)(y_i - y_j) < 0\). Hence, \( \eta \) is defined as:

\[ \eta = \frac{\text{(# of concordant pairs)} - \text{(# of discordant pairs)}}{\frac{1}{2}n(n-1)} \]

Note that \( \eta \) must be in \([-1, 1]\). In particular, if \( Y \) is rigorously increasing monotone with respect to \( X \), \( \eta = 1 \), whereas if \( Y \) is rigorously decreasing monotone with respect to \( X \), then \( \eta = -1 \).