Geo-aggregation of Wikipedia page views: Maximizing geographic granularity while preserving privacy

Reid Priedhorsky, Geoffrey Fairchild, Sara Del Valle
{reidpr,gfairchild,sdelvall}@lanl.gov
Los Alamos National Laboratory

Proposal and request for feedback
LA-UR 15-20145
Draft 2 — January 6, 2015

1 Motivation

Research shows that analyzing internet traces of health-related activity, such as search queries and social media messages, is an effective means of monitoring the spread of disease and has significant promise to tackle problems that traditional disease monitoring tools cannot. The most well-known example of this is Google Flu Trends.[1]

Recently, two teams have extended these techniques to use the globally-aggregated Wikipedia page view logs currently available,[2] with good results.[3,4] There is even promising evidence that disease forecasting, not simply monitoring, is possible as well using these data. Further, these page view logs are, to our knowledge, the only freely available data source in this class, meaning that the science and utility they can support is significantly greater than proprietary alternatives such as Google queries or Twitter messages.

Effective disease monitoring is fundamentally geographic and requires geo-located data sources. Both of the above efforts infer geography at the country level from the wiki language (e.g., views of the Thai Wikipedia are assumed to come from Thailand), whether implicitly or explicitly. Wikimedia traffic statistics generally support this technique.[5] However, it has several problems that make it unsuitable for operational use, including:

1. It cannot be applied at geographic granularity finer than the country level. However, in many countries, disease monitoring must be carried out at the state or metro-area level in order to be effective.

2. Many or most countries can't be covered at all. For example, this approach is not feasible for Chile, which comprises only 4.4% of Spanish-language requests and whose signal is swamped by other Spanish-speaking countries.

[1] http://www.google.org/flutrends/
[2] http://dumps.wikimedia.org/other/pagecounts-raw/
[3] McIver & Brownstein (http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003581).
[4] Generous et al. (http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003892).
[5] http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm

3. It is not dynamic. For example, an outbreak would change viewing patterns for a specific set of articles, not the whole language, potentially yielding misleading results.

In short, the current global aggregation of Wikipedia page view is unsuitable for an operational disease monitoring system. There will be no "Wikipedia Flu Trends" unless page view data are aggregated at a finer geographic scale. However, this aggregation must be performed under strict limits that protect the privacy of Wikipedia readers and editors, which is the subject of this document.

2 Privacy concerns

Were privacy not a concern, we could simply publish the web server logs unmodified. This is obviously not true, first because they contain information that is inherently private and second because they contain information that is private when linked to other information that is already public. The following information about users is of concern:

- *Identifiers*
  - Real name or other real-world identifiers
  - Wikipedia username
  - IP address
- *Sensitive attributes*
  - Geographic location

Specifically, we want to prevent two classes of disclosure:

1. **Reading activity.** Linking of reading activity to any identifiers, either online (Wikipedia username) or offline (real name or IP address).
2. **Location.** Linking of location to any identifier. The sensitivity of this varies. For example, relatively few people would object to their country being disclosed, but more would object to metro area-level disclosure.

In evaluating the privacy of our geo-aggregation method, it is important to select a quantitative metric in order to make precise recommendations. There are a variety of such metrics. These include:[6]

- ***k*-anonymity**: Any given individual resides in an *equivalence class* of $k-1$ others. For example, a given editor can be linked to no fewer than k candidate reading histories. (Note that this example interpretation of k may differ from others in this proposal.)
- **ℓ-diversity**: Any given individual resides in an equivalence class with at least ℓ "well-represented" values of each sensitive attribute. The notion of well-representedness is complex and often ill-defined. A plausible though imprecise example in our case is that any given editor can be linked to no fewer than ℓ locations.
- ***t*-closeness**: Any given individual resides in an equivalence class whose distribution of sensitive attribute values is within t of the global distribution. For example, the probability distribution of an editor's location differs from the location distribution of all Wikipedia editors by no more than t.

[6] Li, Li, & Venkatasubramanian (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4221659).

To select a metric, we need to show that it both is easily understandable (because we will need to argue for our method's reliability in public) and provides the necessary privacy confidence. k-anonymity satisfies the former, and we show below in the discussion of the method that it satisfies the latter as well under the conditions we propose.

3 Method

Our goal is to design a new, geographically finer page view data stream that preserves reader privacy and can be made available to the public. We do so by dynamically limiting geographic granularity and by reducing temporal granularity from hourly to daily. (Technical implementation details are out of scope of this proposal.)

We use the following example:

- Alice lives in Albuquerque, New Mexico and visits the articles "Influenza", "Chills", and "Fever".
- Sam lives in Santa Fe, New Mexico and visits the articles "Influenza", "Chills", and "Chile"
- Carol lives in Calgary, Canada and visits the articles "Influenza", "Fever", and "Hockey".

3.1 Step 1: Build a geographic page view tree

We first construct a geographic tree with four levels: global, nation, province,[7] metro. Both provinces and metropolitan areas descend from the nation level; this is because some metropolitan areas cross provincial boundaries. An excerpt is as follows (the full hierarchy will contain a few thousand nodes):

- Earth
  - Canada
    provinces
    * Alberta
    * Quebec
    metros
    * Calgary
    * Montréal
  - Mexico
  - United States
    provinces
    * Alabama
    * New Mexico
    metros
    * Albuquerque
    * San Francisco
    * Santa Fe

[7] More precisely, "first-level subnational administrative division". Other examples include *department*, *prefecture*, and *state*. We use the common term *province* as a shorthand.

These data can be obtained from Natural Earth,[8] the U.S. Census, and other sources.

3.2 Step 2: Populate the tree

Each article has its own tree; each node keeps a running count of requests for that article from that location, as follows. Every hour, each request recorded in the raw web server is analyzed to extract the article (including language) and fine-grained location (derived from the IP address). For example, the views in the example above would result in the following trees:

"Influenza":

- Earth: 3
  - Canada: 1
    provinces
    - Alberta: 1
    - Quebec: 0

    metros
    - Calgary: 1
    - Montréal: 0
  - Mexico: 0
  - United States: 2
    provinces
    - Alabama: 0
    - New Mexico: 2

    metros
    - Albuquerque: 1
    - San Francisco: 0
    - Santa Fe: 1

"Chills" (this and the remaining examples omit nodes with count zero and the province/metro tags for brevity):

- Earth: 2
  - United States: 2
    - New Mexico: 2
    - Albuquerque: 1
    - Santa Fe: 1

"Fever":

- Earth: 2
  - Canada: 1
    - Alberta: 1
    - Calgary: 1
  - United States: 1
    - New Mexico: 1
    - Albuquerque: 1

"Chile":

[8] http://www.naturalearthdata.com

- Earth: 1
  - United States: 1
    - New Mexico: 1
    - Santa Fe: 1

"Hockey":

- Earth: 1
  - Canada: 1
    - Alberta: 1
    - Calgary: 1

3.3 Step 3: Prune and save the tree

Each day, each article's tree is pruned for privacy and then saved. Then, each article starts over with a new tree.

The pruning algorithm for some k is:

for each level, from bottom to top:
  for each node, from smallest to largest count:
    if the node's count or the total count removed so far is less than k:
      remove the node

This removes from the article's tree all locations with less than k visits and all nodes whose count could be inferred to be less than k. The pruned tree is the one that is saved and published. In short, popular articles will have detailed geographic view breakdowns, while less popular articles will have less detailed breakdowns.

For example, the result of pruning the "Influenza" tree with $k = 0$ for the global level and $k = 2$ for all other levels is:[9]

- Earth: 3
  - United States: 2
    - New Mexico: 2

And for "Hockey":

- Earth: 1

Note that while the above is written to do all nodes at the same time, it could also be arranged to process each node according to a relevant time zone. This would yield daily time series in phase with local external data (e.g., daily disease incidence counts from traditional monitoring).

3.4 Privacy analysis

In this section, we analyze the effect of the above procedure on user privacy. We do so by analyzing how the most easily connected identifier, a Wikipedia username, can be associated with reading histories (i.e., visits to articles other than the one edited) and locations.

[9] The privacy threshold can be as heterogeneous as desired. For example, some countries could be made more privacy-conservative than others.

Two main attacks on k-anonymity are known:[10]

1. **Homogeneity Attack.** This attack depends on homogeneous values of sensitive attributes; that is, even if a user is indistinguishable from $k-1$ other users, if all k users have the same attribute value, then the attribute value becomes known.

   In the case of reading histories, suppose a user Alice edits Article A and then visits Article B (without editing, so this visit should be private). The edit history tells us that she visited Article A (because one cannot edit without visiting). Suppose further that all users who visited Article A then went on to visit Article B, and the only visitors who visit B came from A. However, we cannot recover Alice's visit to B, because the number of other articles is too large to infer that she visited B also. Therefore, her privacy is protected.

   In the case of locations, more can be inferred. Suppose that all visits to Article A were from Alice's metro area. We know that Alice visited A because she edited it. Therefore, we also know Alice's metro, which may be of concern to her. At least two counter-measures are possible:

   - Set a lower bound l on the number of nodes at a given level. If fewer nodes remain after pruning, prune the rest as well. This limits inference of Alice's location to l choices at any given instance, though observation over time will reduce this limit.
   - Don't geo-code requests associated with editing activity. Then, the link between Alice's edit and the visit that leads to her location is broken.

2. **Background knowledge.** This attack uses external knowledge to make informed guesses about the values of sensitive attributes.

   In the case of reading histories, suppose that an attacker knows that Bob likes cooking (for example, from reading Bob's non-Wikipedia blog). Suppose further that Bob edits the article "Spatula" and then visits the article "Spoon". The attacker can infer that Bob is probably one of the k visitors to "Spoon".

   In the case of locations, suppose that the attacker infers that Bob is a native English speaker (perhaps from the same source). Suppose also that visits to the "Spatula" article came from just three metro areas: Tokyo, Shanghai, and New York City. The attacker can infer that Bob is probably in New York City.

   Again, breaking the link between editing and other visits will be helpful.

4 Opt out

Logged-in users can opt out. Requests associated with these users are not geocoded and are not added to any counts other than the global count.

Three possible opt-out procedures:

1. **Active opt-out.**

[10] http://en.wikipedia.org/wiki/L-diversity

- All logged in users are shown a banner during the first n visits after the technique goes live (accounts created after this will be shown the banner immediately). This explains what is going on and how to opt out.
- During this period, the user is assumed to have opted out.
- If the user explicitly opts out, they stay opted out forever (or until they twiddle the bit back).
- If the user takes no action, they are opted in after the n visits expire.

2. **Active opt-in.** Same as above, but the default for logged-in users is no geocoding; they must opt in to be included. Anonymous users stay geo-coded. This would probably work find for disease tracking purposes because the 20 million named accounts[11] (the majority of which are likely abandoned) is a quite small compared to the total number of readers.

3. **Blanket opt-out.** All logged-in users are excluded from location analysis, and there is no setting to reverse this.

4. **Do-Not-Track.** Don't geo-code requests with the Do-Not-Track header. However, many individuals turn this on to avoid advertising and would be fine with our geographic aggregation of visit counts.

We propose using option 3. This is simple to implement and understand and prevents any inference of user account data based on reading activity. Combined with the tree pruning method above, this limits inferrable information to:

- k or more anonymous readers visited an article from a given location.
- Locations associated with IP addresses of the subset of anonymous readers who also edited. However, these locations are already available to the public because they can be looked up using IP addresses published in article editing logs.

5 Human subjects review

We will include the LANL Human Subjects Research Review Board (our IRB) in the design of this privacy plan. In principle, we could leave this review until after the plan was implemented and we were beginning research, but we feel that this review is important to have earlier. (From a logistical perspective, we also need to make sure that the board's concerns are satisfied at a stage when they can be rectified.)

We expect that other institutions' IRBs will examine the plan later as well when other research projects arise.

[11] http://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Number_of_editors