S3. Inferring Subscriber Locations

The call data reports the location of the cell tower used at the start and end of each

call. From the sequence of cell towers used, it is possible to infer an individual’s location.

At any point during a transaction, a mobile phone handset sends packets of information to

one cellular tower, using electromagnetic waves. This tower routes these packets to the rest

of the network using either fiber optic cables or a different electromagnetic frequency; the

packet is sent to a tower near the receiver and ultimately delivered to the receiver’s handset.

Handsets tend to transmit information to a close, unobstructed tower, so that the tower

used represents an approximation to the individual’s location at that point in time. Calls

can bounce between towers due to call traffic, variation in the weather, if a tower is down,

or if the handset is in motion. The maximum technical range of a GSM tower is 35 km, but

in areas of higher tower density the range is reduced to lower interference.

There is a literature on inferring a subscriber’s location based on usage traces (Gonzalez

et al., 2008; Isaacman et al., 2010, 2011; Blumenstock et al., 2011). My settings differs from

these papers in two ways: the tower network was rapidly expanding, and usage is sparse. I

implement a modified version of the ‘important places’ algorithm as detailed by Isaacman

et al. (2011), which for each user identifies one or more important places where they spend

time. The paper finds that the identified places were within 3 miles of reported places for

88% of a small validation sample of users in the U.S., with a median error of 0.9 miles. I

have modified the algorithm to improve performance in rural areas.

To find the important places for individual i, the algorithm proceeds as follows:

(1) The towers that i has ever used, Xi, are sorted by the number of days i used that tower,

dix

(2) The most used tower forms the start of a new cluster, located at that tower’s location.

(3) If the next most used tower falls within a distance threshold of the cluster, it is added to

that cluster, and the cluster’s location moves to its new centroid (weighted by the days each

tower is used). If the tower does not fall within the threshold, it forms a new cluster. The

original paper uses a fixed threshold of 1 mile, with which they obtain good results in an

urban setting. To allow for good performance in urban and rural areas (high and low tower

densities), I compute an adaptive threshold specific to each tower related to the density

of towers nearby. In considering the distance from tower x to a cluster, I use a threshold

equal to the distance from x to the 9th most distant tower as of May 2009. This adaptive

threshold allows the algorithm to smoothly incorporate a large radius of spatial information

in rural areas and a narrow radius in urban areas.

(4) The previous step is repeated for each tower: if the nearest cluster is within this tower’s

threshold, the tower is assigned to that cluster and that cluster’s centroid is updated; if the

nearest cluster is further away, the tower is assigned to a new cluster.

(5) After all towers have been placed in clusters, each cluster is ranked by the combined days

that the individual made calls from that cluster (counting each day only once if transactions

were made on multiple towers within that same cluster).

This algorithm has advantages for this setting: it uses the full panel of data, which improves

precision when transactions are sparse, and works well with an expanding network: estimates

simply become more precise as tower density increases.7