==========

IMC'09 Review #4A

Updated Wednesday 10 Jun 2009 2:39:30am CEST

Paper #4: Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

---

Overall merit: 3. Top 25% but not top 10% of submitted papers

Reviewer qualification: 3. I know the material, but am not an expert

Novelty: 4. New contribution

===== Paper summary =====

The paper analyzes a dataset of data traffic from more than 280k mobile phone users. From that dataset the paper approximates mobility pattern (to the precision of base station), classifies application usage by mobile and non-mobile users, identifies "hotspots", and looks for users accessing common locations, either at the same time or different.

===== Exciting =====

The paper has a large, real dataset evaluating what people really do with mobile phones, there's a lot of potential in this paper.

===== Run-of-the-mill =====

Nothing.

===== Limitations =====

Some of the clustering seems questionable, particularly the regional analysis (sec. 5).

===== Comments to author =====

This seems like a very nice paper with a very interesting dataset.

Some questions:

- can you comment explicitly in the paper on dataset availability?

(*) Unfortunately the dataset cannot be made publicly available. As noted in the reviewer comment below, such a dataset cannot fully preserve user privacy, thus releasing it can prove perilous. We mention this in the paper.

"We note however, as observed in previous work [15], that one can still infer private user characteristics from such data. This is one of the reasons why the dataset used for this paper cannot be made publicly

available."

- the claim in section 2.1 that the data "completely preserves user privacy" seems in accurate given recent work mapping home and work locations to user identity. See "On the Anonymity of Home/Work Location Pairs" by Golle and Partridge.

(*) We agree with the reviewer that considering such data one can infer certain user characteristics. We have consequently toned down the above comment (by removing the word "completely") and currently cite the mentioned paper.

- It is not obvious if your three definitions in section 3.1.1 are overlapping or not. Can you please clarify that? Also, the definitions are very awkward and hard to follow. The English is basically math written out with words. Can you clarify them?

Also, would disappearance be easier to recast as movement to the location "nowhere", allowing you to eliminate one rule?

(*) The definitions of the rules are explained in the paper as follows:

"First, note that a user who is accessing the mobile data network from a certain location (base-station) has the following three possible exit states: (1) user moves to a new location either while staying connected via the same session (intra-session movement) or logs-off and logs back again via a new session (inter-session movement); (2) user stays at the same location (intra- or inter-session stationary) and; (3) user switches-off his mobile device (disappear) and does not re-appear for some time."

The definitions are consequently constructed using the above simple observations. Further, since a user can be in only one of the above circumstances the definitions are not overlapping. In fact, when constructing the rules we wanted to preserve the above mentioned semantics.

We have added in the definition of the movement rule (Definition 1 in the paper) the constraint that the movement location cannot be "nowhere" to prevent misunderstandings.

- Several of the conclusions seems obvious. For example, in Sec. 3.1.4: "Interestingly more than 70% of mobile users visit at least one common location on every single day...". Doesn't that mean: everyone with a mobile device has a home?

(*) This particular conclusion was drawn in relation to previous work that deals with the high predictability of human movement and was not strongly

emphasized in our paper.

- Some of the clustering seems questionable, particularly the regional analysis (sec. 5): You group the data into k clusters there, for 1<=k<=5. Is 5 a meaningful number of clusters for 280k users? I think you could get arbitrary results in this section by varying k, and I'm not convinced these arbitrary clusters are particularly meaningful.

For example, you draw the inference in section 5.1 that the number of unique people sharing the same interests is larger in regions with more people. This conclusion seems obvious. But also, it seems nearly useless, because the definition of "interest" is very broad (if I understand, anyone running music, or mail, or any trading app, anywhere in downtown, has the same interest as I do by the paper's definition).

(*) First, by clustering you can indeed get a variable number of clusters by varying k as the reviewer observes. We did exactly that by varying k until 32. We obtained the 5 big clusters we presented and focused on and the rest being small ones comprising a small number of locations and users. This is mentioned in the paper: "We run the multi-partitioning algorithm 1 with different values for the number of desired clusters k and across multiple runs of the algorithm, we always identified the same five significant regions."

Second, the inference that the reviewer suggests is obvious was never meant to be presented alone. We presented this result in contrast to the second result regarding the frequency of interactions in hotspots. Regarding the definition of interest, it is indeed broad and as we stated in the paper it is meant to be regarded as an upper-bound for such services.

Minor comments:

Table 1: please define BSID. And Is "avg." here mean or median?

(*) We have fixed this."avg." means mean.

Table 2: your footnote on music is labeled "2", but appears as "3" at the bottom of the page.

(*) We have fixed this.

Figure 2: there seems no real difference between these values.

(*) This is explained in the text and is in line with the results observed in the reference:
"M. Gonzalez and C. Hidalgo and A. Barabasi. Understanding Individual Human Mobility Patterns."

Figure 4: wouldn't this figure be clearer as a stacked bar chart?

(*) Presenting this figure as a stacked bar chart would not be particularly useful since the number of patterns required is 6 and the individual values are close together. Note however that the trends presented in the text can be easily spotted in the figure.

==========

IMC'09 Review #4B
Updated Thursday 11 Jun 2009 11:09:12am CEST

Paper #4: Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

---

Overall merit: 3. Top 25% but not top 10% of submitted papers
Reviewer qualification: 4. I know a lot about this area
Novelty: 4. New contribution

===== Paper summary =====

This work mines data comm data from 280,000 users of a 3G network over a week and reports on their top intentions and mobility patterns.

===== Exciting =====

Correlation of mobility patterns to Internet access interests is interesting.

===== Run-of-the-mill =====

The paper does not raises why and provide supporting evaluation. Strictly data-driven to the point that results are either predictable and boring or matter-of-factly without knowing why we see that behavior.

===== Comments to author =====

I would like to point out one distinction that is not clear in the paper but would have an impact in the overall evaluation. Can you distinguish between cell phone and modem uses? I find it rather strange that the most popular application at home is music. People have radios, TV music channels, audio systems and MP3 players at home and I see no reason to listen to music over 3G network at home. But if the person is streaming over the network, maybe. The device of choice should have a big impact on the overall user data comm pattern and be made clear. Certain devices double as navigation tools. Device-dependent features can have a bias on the outcome. If the data does not have this information, the paper should be explicit about the limitation in the interpretation of the analysis.

(*) We agree with the reviewer that such device information can be very useful in our analysis. Unfortunately our dataset does not contain such information. We acknowledge this in the paper:

"One limitation of our study lies in the fact that our trace does not contain device type information. Indeed, certain devices have characteristics which make them attractive for a specific purpose, for example they can be easily used as navigation tools or for sending e-mails. Such device-dependent features can have a bias on our analysis. Also, recently, mobile service providers have started commercializing modems that can be used with personal computers such as laptops. Users can use these devices to connect to the Internet from anywhere within the cellular network. One concern is that these devices do not constrain the user (in terms of application accesses) in the way a limited resource platform such as a mobile phone might. Such modem devices exist in the network that we have analyzed, although in a relatively small proportion compared to the total devices; they number a few hundred as informed by the provider."

Social network-based LBS should start with social network, while the data of this work contains no such information. Lack of this information about human network make it very hard to evaluate the feasibliity of LBS services. People of common interests at the same location could be introduced, but having the same interest is not as important as the intention to participate in such a service. Once joining sucha recomm service, people can choose to be introduced to those with opposite interests, or solely based on the look. If I am at Kings Cross in London, I'm sure there'll be at least a hundred of people who will share some common interests with me. Should I be interested? I find the analysis in Section 5 too limited in the sense that they analyze and evaluate synthetic questions of little relevance in human social behavior.

(*) First, we agree with the reviewer that carrying out such a study would be interesting. However, from the data at our disposal we cannot estimate if the people carrying the devices actually have a connection in real life. In the author's opinion such a study would be hard to conduct at a large scale.

Second, intention to participate in such a service is hard to estimate. Some users might have no problem in disclosing location information and entering such a service while others might prove conservative and use the usual web services. As this relates to the user, it cannot be evaluated by network means. Also, as mentioned in the paper, the analysis in section 5 should be considered as an upper bound in terms of interest for such services.

"SLAW: A Mobility Model for Human Walks" offers a mobility model and also explains the grouping of users in mobility. Their work is relevant to the user affinity analysis in certain locations and should be tried.

(*) The reference given by the reviewer offers a mobility model that manages to generate realistic human mobility traces. Although it would be interesting to validate such a model we find it beyond the scope of this paper.
We cite the above paper in the related work section.

_ _ _ _ _

==========

IMC'09 Review #4C

Updated Thursday 16 Jul 2009 12:14:06pm CEST

Paper #4: Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

---

Overall merit: 4. Top 10% but not top 5% of submitted papers

Reviewer qualification: 4. I know a lot about this area

Novelty: 4. New contribution

===== Paper summary =====

The paper describes an analysis of the 3G access patterns of 280K users in a large metropolitan area. Thanks to the data, which includes handovers between cells, the authors can associated user's mobility to web access patterns. They classify the web pages visited by the users into 15 different groups ranging from dating to video in order to study the interplay between location, time of the day and web usage.

The authors present many findings that can be extremely useful for the fields of content distribution networks and delay tolerant networks. Although some of the observations were to be expected they were lacking empirical evidence. Some of the observations, however, are not so trivial and they can have a strong impact on system design.

===== Exciting =====

+ Analysis of a novel data-set that combines location with web access on a metropolitan 3G network.
+ Relevance of the findings of the paper to to other fields such as DTN's and CDN's.
+ The paper contains many interesting observations. Some of them were to be expected although no empirical evidence existed prior to this work. Other observations are not so trivial and can have a strong influence on content distribution strategies for cellular networks.

===== Run-of-the-mill =====

The paper lacks a proper study on the temporal patterns of web access. Section 4.1 is not convincing enough. I would have expected an analysis on the application access by time without accounting for hotspots, which could be introducing confounders.

===== Comments to author =====

This is an excellent paper that could be on the top 5% if it wasn't for some issues enumerated below (specially #5)

1) Is the common location that 70% of users return to their home? You should be able to infer that from the base station location as well as with density. In any case, I was expecting the highest common location to be larger a higher probability, 30% of users do not return to their home during one day?

(*) Our data regarding human movement is sampled in the sense that users connect and disconnect from the data network as needed. Therefore we cannot draw the above conclusion simply because users connect to the data service at their will.
---

2) The results of the inter-session movement and stationary depicted in Figure 3 are kind of surprising. Although the daily patterns (night and day) appear, the confidence probability of movement at its highest is only about 3 times larger that in the middle of the night. I would not expect such a small difference in the confidence probability. This point should be further discussed in the paper, otherwise it might arise some suspicions on the bias of the data-set.

(*) The figure depicts the confidence probability which is by definition normalized to the support at that corresponding hour. One can note also the difference in support between the hours of night and day. So actually, the difference in the normalizing support and in the confidence probability actually account for a larger difference in the confidence values between the hours of night and day. This is explained in the text:

"Note that the confidence probability is by definition normalized by the values of the support at the considered time interval. Therefore the ratio between the values of confidence for the hours of the day and the hours of night is larger than the ratio between the values in the confidence probability for the hours of the day and the hours of the night."
---

3) There are other hypothesis besides bandwidth and battery consumption to why music is so high in the comfort zone. As a matter of fact bandwidth should not be a factor since the data-set only accounts for 3G networks and the residential coverage is not as extensive as downtown coverage. It would be interesting to know whether the users are accessing music from streaming (pandora) or not. Users could be building a a playlist for the next day which requires more attention and therefore it has to be done in the comfort zone. You do not take into consideration what is an evident distinction between the applications; whether the applications is work or leisure related. If you factor this aspect most of the observations are clear cut. Naturally work related applications are more frequent outside comfort zone and during work hours, while leisure such as music or dating operate within the comfort zone.

(*) We agree with the reviewer that this would be an interesting direction to purse. However validating such observations (building a playlist or not) can be particularly hard by network means. The same applies for distinguishing between work and personal use of an application.
---

4) The classification of URL into interest is completely ad-hoc, which is fine. However, more details should have been included in the paper. What percentage of the URL are not classified in any group? What is the overlapping between URL's? I see quite a few holes in the keyword selections e.g. loopt, youtube, twitter.

(*) The classification covers 94% of the URLs. The rules are applied in

order and not in parallel such that there is no overlap in the classification. We did not observe a large number of accesses to the services mentioned by the reviewer.
---

5) The biggest concern of the paper is whether application access is mobility or time based. This question arises in section 3.2.1. and it is somehow addressed in section 4.1. However, section 4.1 introduces the hotspots that can confound the results. You should have analyzed the temporal application access to fully discard temporal correlations. Yet it is true that from Fig. 8 one can see that location is driving the application access this could be done due to a different macro-behaviour, for instance people commuting.

(*) The authors haven't discarded completely the temporal patterns in application access. It is clear that temporal patterns play a role in the way users access content. Our conclusion regarding this is that besides temporal patterns, also location plays a role in application access. Note that in section 4.1 the conclusion drawn summarizes the above points: "Hence, we conclude that time-of-day does not dominantly affect the accesses at hotspots."
---

6) In page 4 there are 2 footnotes labelled 3.

(*) We have fixed this.
---

7) The last issue is on section 5.1. I concur that the empirical findings of this paper can guide the design of better content distribution systems that leverage random encounters. However, the authors are way too optimistic on the drawn numbers. The high number of interactions -- or encounters -- is misleading. The fact that they are in the same cell do not imply a proximity contact by which data could be exchanged via wifi or bluetooth. Furthermore, doing the transfers at the cell level could be too cumbersome for the operator. Another aspect by which the results of this section are not conclusive is that the authors assume that application access correlates with interest. Not everybody who access email is interested in meeting other email users, not all people who listen music are actually interested in the same kind of music, etc. The point of this section is understood and it is one of the many
applications of the author's measurements, however, it has to be contextualized better.

(*) We have addressed the first point in the introduction. Indeed the location information we have is coarse-grained as noted in the introduction and we cannot infer from this whether users in the same cell are also within wifi or bluetooth distance. We cannot speculate on doing transfers at the cell level, yet we note that most operators offer shops where users can download music/ringtones/games over the network.
Second, intention to participate in such a service is hard to estimate. Some users might have no problem in disclosing location information and entering such a service while others might prove conservative and use the usual web services. As this relates to the user, it cannot be evaluated by network means. Also, as mentioned in the paper, the analysis in section 5 should be considered as an upper bound in terms of interest for such services.

_ _ _ _ _

==========

IMC'09 Review #4D

Updated Friday 3 Jul 2009 12:53:23pm CEST

Paper #4: Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

---

Overall merit: 2. Top 50% but not top 25% of submitted papers

Reviewer qualification: 5. I am an expert on this topic

Novelty: 3. Incremental improvement

===== Paper summary =====

The paper uses a data log of 280,000 users of a 3G mobile network in a large metropolitan area to characterise the relationship between people's interests and mobility properties. Their analysis reveals that (i)people's movement patterns are correlated with the applications they access, (ii)location affects the applications accessed by users, (iii)and the number of serendipitous meetings between users of similar cyber interest is larger in regions with higher density of hotpost. (i) and (ii) actually are talking about the same thing, and (iii) is as expected.

===== Exciting =====

It is an interesting topic to study human mobility and the application they access during different mobility mode.

===== Run-of-the-mill =====

The definition of the different rules are a bit confusing, and it is easier to express by a few world instead of puting something confusing.s

===== Limitations =====

There are several major problem that I would propose to put the paper into a weak reject category

1. It is not sure that the users are using UMTS or other 3G USB sticks with their laptops or they are accessing the 3G network on their mobile phones. This definitely make large different for the results and analysis of the paper. If the users are using UMTS USB stickt, the behaviour during the so called hotspot/comfort zones would be very similar to normal internet usage. For example in Europe or some developed city in Asia, people may only have a mobile UMTS USB stick for even home use instead of subscribing to a broadband. And of course, when the people at home, they will use the network to listen to music and doing social network. And when you are traveling, you will probably just check emails since you will be too busy to do music listening stuffs. The authors should make it clear on this aspect, and this is one major problem. I am sure there are some laptop or desktop users among them, we can see from from page 9 "because users can have more than one application affilic
ation, the sumer of normalized affiliations does not equal to one". On mobile phone, usually you can run one

of such application at the same time.

(*) This comment was answered above when considering the comments of a previous reviewer.
Further, even mobile phone users can have multiple application affiliations. When connected to the data network with a mobile phone one can access any kind of service even at the same time depending on the capabilities of the phone.
---

2. The authors is kind of missleading by giving a large number of 280,000 users. How many of them are active users? Looking at the support on Figure 3, the maximum support is less than 37,500. A futher evidences of this unreliable of this number is that on page 9, they identify only 23 day hotspots, 28 noon hotspots, and 8 evening hotspot. Is that true that all thease 280,000 users only have 23 day hotspots? This is not convencing, especially many areas are covered by multiple base station. I think if the authors want to draw a more scientific conlusion, they should extract the activit users instead of giving a big number.

(*) First, the number we gave, 280,000 users is indeed the number of active users. All of the 280,000 unique users have used the data network at least once during the seven day interval.
Second, regarding the hotspots we have found the emphasis was not on their number but on their temporal characteristics (regions becoming extremely populous during specific parts of the day).
---

3. Also there is a major problem for classifying mobility. Currently, the authors classify movememnt as change of cellular tower ID. It is well know that in cellualr network, the cellular phone will keep associate with different cellular towers nearby even the phones are in stationary. This is normal for a place covered by multiple cellular base stations. This can explain why the author observed 84% of session spend less than 10 seconds in motion (page 6). Many of this kind of motion can be because of the celluarl phone or laptop keeping swinging among different base stations at the same area. If the authors cannot filter out this effect, the analysis of mobility cannot produce a scientific conclusion.

(*) The reviewer is incorrect. The cellular phone keeps associated with a single cellular tower at a point in time. This is the tower that is currently serving him. The phone does indeed measure power levels from other towers (by monitoring the signals coming from these towers) and if the power level of the tower that is currently serving him drops under a threshold, it associates with a tower that offers a better signal. Even in a location with multiple base stations, the situation is as described above, the phone is served by a single base station and monitors the signals of several so that it can switch.

The 84% of sessions that spend less than 10 seconds in motion are mostly stationary sessions since less than 10 seconds in motion also includes 0 seconds in motion. This is explained further in the text:
"Once again, about 84% of sessions are completely or 100% sedentary while 6% of sessions are completely mobile."

We have also made the text clearer in the corresponding section:

"While a majority 84% of sessions stay stationary (that is stay within one base-station), and 6% move for 15 minutes or less, the 99%-ile is 3.5 hours and the maximum time that a session spends in motion is one day."

===== Comments to author =====

Minor problem:
1. In the conclusion, the author said" in this paper we conducted, to the best of our knowledge, the first large-scale experital study", this is missleading. They did not conduct the experiment, but got the data from operator.

(*) The reviewer is correct. We have removed the word "experimental".

2.citation [10,16] are not using GPS information as the authors mentioned. and also Levy flight is not reandom models.

(*) The reviewer is partially correct. Citation 10 does not use GPS information. It uses multiple datasets from which it extracts device contact information. Citation 17 uses multiple datasets of which one is based on GPS information. We have fixed the citations accordingly.
Levy flights are random in nature: random walks with increments distributed on a heavy-tailed pdf.
---

3. I am not sure the bipartite graph G they constructed in page 11 is correct. There are 281,394 users, and 1,196 locations, and as they said if a users never visted a location, the weight of the edge is 0 but there should be still an edge. Then the total number of edge 936,280 should be wrong.

(*) The graph is indeed correct. If the weight is 0 in the adjacency matrix it means there is no edge and so the total number of edges we give is correct.
---

4. some conclusions are a bit obvious, for example "suggesting that users regularly revisit their useral location", "the probability of meeting different people is larger in a more populated region"

(*) Regarding the obviousness of this particular conclusion we have addressed this in a previous reviewer comment.
---

5. It is better to plot Figure 5 as a CDF.

(*) The main point of Figure 5 is that the top 3 locations account for most of user presence and this can be seen clearly from the present figure.

==========

Comment

Paper #4: Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network

---

Among the issues and comments that arose while discussing you paper at the PC meeting are the following:

* What kind are the user-terminals? What devices were used? (Handhelds, smart-phones, UMTS-sticks, vendor, etc.)

(*) This comment was answered above when considering the comments of previous reviewers. We have addressed this issue by updating the text at the end of Section 2.3 in the paper.
