
==========
SIGCOMM 2009 Review #52A
Updated Monday 2 Mar 2009 5:16:01am EST

Paper #52: The Genesis of Location-Aware Mobile Social Networking
---

Overall score: 2. Likely reject. Weaknesses outweigh strengths.
Reviewer confidence: 3. Good confidence
Technical merit: 2. Fair
Novelty: 3. Incremental improvement
Relevance/community interest: 2. Low
Longevity: 3. Not important now, possibly short lifetime in the future

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

This paper presents a measurement study of mobile application usage patterns for a population of 280,000 users. The authors study both mobility patterns and application usage characteristics for this population and draw inferences on the possibilities of mobile social networking between various users.

Strengths: The measurement is over a large data set, hence the authors can claim high confidence in them.

Weaknesses: The authors are trying to judge the feasibility of mobile social networking using cell towers as an indication of location. This is too coarse a granularity to make any inferences on social interactions. Moreover, a number of the results on mobility patterns are not novel or are unsurprising.

===== Comments for authors =====

While the measurement study is interesting, drawing inferences on social networking possibilities using this data is a little dangerous. Coarse-grained proximity, without the use of GPS or user input on precise location, does not imply compatibility. Moreover, while the authors do make conjectures on how this data could enable more social networking applications, no concrete efforst to apply the results to relevant applications has been shown.

The title of the paper is a little misleading. Most of the paper is dedicated to low-level measurements on mobility, application usage, and hot-spots. There is very little on location-aware, mobile social networking.
A suggestion would be to either concentrate mainly on the measurement and inference angle of the paper or validate how these measurements aid the design of location aware social networking.

Section-by-section comments:

Sec. 2: The authors do not describe where exactly they collected all this data.
Sec 2.2: Isn't Figure 2 a cumulative distribution function?

Sec 2.3: Why is music considered a social networking app? For all you know, the user could be logged on to itunes or a radio service.
Sec 3: The varying values of \delta make it difficult to understand its significance in various contexts.
Sec 3.2.2: What does "potentially counter-intuitive" mean?
Sec 4 and others: Several inferences that the authors make are stated almost as facts. The language should make it clear that these are just inferences and have not been validated (if that is indeed so).
Sec 4.1: Not clear how the authors infer that "time-of-day does not affect the acceses at hotspots" from Figure 9(b).

Detailed comments

==========

SIGCOMM 2009 Review #52B
Updated Friday 17 Apr 2009 6:48:02am EDT

Paper #52: The Genesis of Location-Aware Mobile Social Networking

---

Overall score: 2. Likely reject. Weaknesses outweigh strengths.
Reviewer confidence: 3. Good confidence
Technical merit: 2. Fair
Novelty: 4. New contribution
Relevance/community interest: 5. Exciting
Longevity: 2. Important now, short lifetime

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

This paper applies association rule mining and traces analysis of mobile web access in order to characterize how mobility relates with network service access. The methodology follows classical data mining techniques, the application to this domain is new to the best of my knowledge; combined with a unique data sets, the paper does provide several interesting perspectives, in particular on possibility of ad-hoc location based services.

Strength:
- exciting topic: looking at new data sets to
- novelty: the paper introduces new data mining approach to this field,

Weaknesses:
- a bit superficial: it generally claims to provide an "explanation" whereas it actually shows one result together with a basic intuition. It rarely presents a cross validation of the actual claim.
- The paper is purely descriptive. It fails a bit short in "networking" result; in fact it does characterize notions like clusters, hot spots, users interactions, that seems very relevant for future networks, but it

does not exploit them in any way. So the actual importance of this notion is not validated by a network experiment (or simulations) to demonstrate their relevance.

Overall, this paper is an interesting read because it deals with a very interesting topic (location aware nomadic services) with unprecedent data sets and methodology. Perhaps the most interesting contribution is to show that location seems a primary factor for nomadic usage (more important than time of the day).
On the other hand, technically none of these results are very strong (figures from association rule learning, statistics from nomadic service usage, spectral clustering without any details on the input etc.). They are neither easy to interpret or map to a networking problem per se, nor compared with previous findings. A lot of claims are somewhat exagerrated (or rather obvious), which weakens a little the contribution of the paper in terms of actual results.

===== Comments for authors =====

I would really wish to put the highest score for this paper, as I think it really address an interesting topic and contains some novelty in the treatment of the data. The "hot-spot: location matters more than time" is the most interesting and sound result in this work. Unfortunately I have to admit that the results leaves a lot to be desired, actually almost all other results. The paper suffers from claiming too many times a big findings by comparing two data points and providing no validation.

I think some of it can be fixed, and the paper is still novel and will generate interesting research, I then propose to accept it in Sigcomm provided room is available and it does not prevent a paper with a very solid contribution to appear.

===
General remarks:

- Overall, the claims made on top of the observation seems very debatable. I would suggest to be more factual, rather than providing conclusions about the result in a vague and imprecise forms.

1) To start with, some claims are completely unappropriate:

p.1 abstract
"(ii) Those who move show highly predictable patterns, yet a person tends to access different applications at different locations"

Firstly, the fact behind the first claim is simply that users spend an important time in their three most frequently used locations (>55%). Is that a "highly predictable pattern", why? Secondly, I see no results backing up the second part of the sentence.

Same comments about
p.1 col.2 "Our analysis
confirms previously reported results on the high pre-
dictability of human movement."
which results? what could contradict them and why did you prove it's not the case. The fact that people follow some skewed frequency of visits is already widely known, is that your main justification?

p.1 abstract

"Our analysis demonstrates
how location-based services can benefit from the knowledge
about human mobility."
Where is that done in this paper? Where is a networking scenario where the knowledge (btw of what?) helps to deliver better location based services. I agree it's possible and that this paper is useful to get to that goal, but this paper does not demonstrate that at all.

p.1 col.2
"our study advances the state-of-the-
art knowledge, e.g., [16], about the basic laws governing
human motion."
This point is vastly ignored in the rest of the paper. True there are some clusters, some preferred locations, but all these notions have been introduced in this context for years, and generally are much better treated than here. I see zero addition to the state of the art when it comes to mobility only. However, I think the nomadic application dimension is new, and should be put upfront.

2) Other claims have somewhat a real ground (one observation taken from the data) but currently they seem exaggerated. More caution would be required. I believe this can be fixed in due time for the paper publication.

- p.2 col.1 "We find that the probability to meet differ-
ent people with the same cyber interests is dominantly
impacted by the number of users sharing the same in-
terests in a given region."
and in p.14 col.1 "The probability to meet differ-
ent people with the same cyber interests is dominantly
impacted by the number of users sharing the same in-
terests in a given region."
What does "interest" mean here? is that an application "affiliation"? This claim seems only justified by a very trivial observation:
p.12 col.2 "This is because the number of users is much
larger in region 4 (82k) than in region 1 (54k). Hence,
even if the user mobility patterns are similar in both
regions, the probability of meeting different people is
larger in a more populated region."

First, this observation has nothing to do with "interest" as it seems to deal only with a region's population. Second, making such a claim ("dominantly impacted") from a comparison between two data points (cluster 1 and 4) looks total bogus: what are the other factors that have been shown not to impact as much? none that I can see.
This point should be either (1) better justified, in particular w.r.t. different interest or simply (2) removed.

- Similarly the second claim (density of hotspots primary factor to explain number of interactions) is a pure intuitive ad-hoc arguments based on two data points (again cluster 1 and 4). It does not provide any universal or general law, it simply seems to fit right with the intuition, but in no way this paper can be considered a proof of this fact.

BTW

"The key reason is
that region 4 is more densely populated with hotspots."
I guess the authors mean region 1.

- Last, the question mentioned in intro and 5.1
p.1 col.2 "how likely is it to meet in our daily lives,
and where, with people who share similar interests in
cyber domain?"
p.12 col.1 "Here, we explore how probable is it, and what deter-
mines the probability, for people who share the same
interests in the cyber domain to meet as part of their
daily lives?"
is not addressed: the paper at no times provide an empirical estimate of the chance to meet someone (or the time to wait to meet one) with similar interest, it simply shows trends.

p.7 col.1 "We explain this phenomenon later in the text,"
Again, you do not explain it, you provide an intuition that is not contradicted by your observation, but there seems to be millions of possible explanation for it, and no test is provided to sort them out.

Note that the "explanation" music is typically is static places is shown via comfort zone, but since music is apparently the dominant application it may be simply that the comfort zone (most used) is precisely the places people use the favorite application (download music).

- p.12 col.1 "Second,
people in the two urban areas, represented by clusters 1
...
the percent is larger than 70% in both cases."
We see 60, 63 and even 69,9 for other clusters, do you think a difference of 0.5 or 7% is that significant with this metric?

3) some points are obscure:

- Fig.2 "Figure 2 shows the complementary distribution func-
tion (CDF) of ..."
I would feel reasonably comfortable with the conclusion of this paragraph, but I am not at all comfortable with the proof. How can you make any claim about representation of a sampled data sets with the use of this data set? Why precisely the fact that human movement remains bounded in general helps? Currently I even think the paper will be stronger without this "validation" that appears to me at best unreadable and at worst flawed.

- Section 5 relies on a division of space (among users or locations?) that is completely ignored in the development. Without any description and sentences like:
p.11 col.2 "We omit
details, yet demonstrate below that spectral clustering
works very well."
it's hard to see the relevance of these results.

- p.3 "there is on average a gap of 6 hours and 11
minutes between two consecutive sessions from the same

user. Moreover, the average gap between the time when
a user accesses the network from a location and then ac-
cesses it from a different location (inter-session move)"
hard to follow, maybe it will be important to define precisely what you call a "session" since it does not seem to be defined with a single location.

Local comments:

- p.5 col.2 "two local peaks" given the relative minor differences with other confidence probabilty do you consider them significant peaks?

- p.6 Fig.4 (a) weird. why not plot CCDF in log-log scale as usually done.

- p.8 Fig.6 Again, why not provide a CDF like Fig.7 the current numbers quoted in the text are actually a CDF one.

- p.8 "We consider that a location is a users home
if we observe a user spending the most time between
10 PM and 6 AM at this location."
I guess the authors mean the location that is the most frequently visited during that time, which is likely but not necessarily the one where most of the time is spent (depends on the distribution).
BTW why not define a home and a work, which may in some cases coincide.

- p.10 How do you explain that mail is low everywhere at all time (as shown in Fig.9 (c)) and yet it appears as dominant in the user base? This looks like a contradiction, unless one time is not covered at all by hotspots (night, but is that relevant?).

- p.11 "Hence, we conclude that time-
of-day does not affect the accesses at hotspots."
rephrase, it looks absurd, of course it does affect, but not as clearly as location of hot spots that's all.

- p.12 "Figure 10 shows the number of time-independent and
-dependent interactions as a function of different loca-
tions in regions 1 and 4."
I do not see the point of this x axis. Can you explain? Besides, the y-axis is not renormalized per nodes, so it looks totally abstract, and comparison among clusters looks absurd.

- p.13 "as predicted by earlier models"
Absurd. These are assumptions, not prediction of a model. How can a model predict its own assumptions?

==========

SIGCOMM 2009 Review #52C
Updated Sunday 29 Mar 2009 7:05:37pm EDT

Paper #52: The Genesis of Location-Aware Mobile Social Networking

---

Overall score: 2. Likely reject. Weaknesses outweigh strengths.
Reviewer confidence: 4. High confidence
Technical merit: 1. Poor
Novelty: 4. New contribution
Relevance/community interest: 4. High
Longevity: 2. Important now, short lifetime

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

This paper is a measurement study of the workload and the mobility patterns of a mobile network with hundreds of thousands of users. The paper poses a series of very interesting questions, such as:

- What is the workload of this mobile population? What type of applications are most popular?

- Is this workload changing depending on the users' locations? Time-of-day?

- How likely are people running similar applications to meet?

Unfortunately, the paper's answers are very unconvincing because problems with the methodology or with the data analysis. The methodology is either not completely presented, or when presented it is unclear. Also, the analysis is confusing. There're basic questions about the data, such as whether this data is WiFi data, 3G data, or data from users of some sort of other wireless technology ????

Overall, the paper poses an interesting set of questions but fails to answer them in a convincing manner. I suspect that a clear description of the methodology and the analysis would make some results appear very unconvincing or even wrong. See the detailed comments for more detailed criticism.

===== Comments for authors =====

Concise questions about the methodology and analysis:

1. Is this WiFi or 3G?

2. Is the data analysis a sample only? Paragraph 1 in Section 2.2 seems to indicate so. If so, what fraction of the total data is this sample?

3. How many locations is data collected from?

4. What is the typical battery lifetime of the users?

5. How large is this area? What's the density of access points? Are there any "dead" regions? Are there regions where users see multiple access

points?

6. Is there a bias in how access points are deployed? Aren't access points already deployed at high traffic places? What sort of bias does this introduce in the results?

7. The distance between people is computed by mapping locations to zipcodes and taking the average distance between the zipcodes (Section 2.2.1). Isn't this introducing a large error?

High-level questions about the paper:

- I don't understand the title. Is the title implying that this is the first paper on "location-aware mobile social networking"? I'm pretty sure I attended at least a workshop session with the same title.

- I can't figure out what the role of the first three paragraphs of the Introduction is. Is it to motivate the need for mobile social? For location-aware? Are there any connections between them and the rest of the paper?

- I also had a hard time understanding the questions raised. What does "application affiliation" mean? Is it workload? Also, the paper poses a question (para 4) to immediately back away from it and pose a different question (also para 4). Very confusing. It's also unclear at this point in the paper what the relationship between "mobility properties" and workload is; they seem quite different.

Detailed comments:

- Section 2.1 -- it seems that the geo information is very coarse-grained, at the granularity of zipcodes. Let's take the postal delivery worker. To me, he's mobile, although he spends *all* his time in one zipcode. I couldn't figure out what mobility means in the context of this paper.

- Section 2.2.1 -- what does "core mobility patterns" mean?

- Section 2.2.1 -- CDF. I think "complementary distribution function" should be "cumulative distribution function".

- I had a hard time understanding what Figure 2 shows? Why isn't there a 0km distance? How can a user move as much over 20mins as over 1day? Also, except the 20-min curve, all other curves have points with x-coordinates of 100km. How could a user move 100km in 40 mins? ?? Also, I don't understand the claim that Figure 2 is "in line" with the findings in reference [12]? Is this about the point that most humans travel short distances most often? Isn't this sort of obvious?

- I think the methodology for keyword-based URL mining is slightly

wrong. There are many URLs I visit with the word "google" and "msn" that are not about search. See "google.com/mail" and "msn.com"? Google and msn serve as portals to a large degree.

- I do not understand what the paper means by mail, trading, and news "comprehensively represent a wide range of behaviors." ??

- I found the Section describing the "Rules" confusing. It's full of Greek symbols, but I think these are relatively simple concepts. Why not use English to describe them intuitively also?

- Figure 3, both graphs, have a "Support" line. Is this described anywhere in the text, other than giving the definition of what "support" means? What do the curves show? Also, I couldn't follow very well what the points of these graphs or section 3.1.2 are.

- Section 3.1.3. 84% of sessions spend less than 10 seconds in motion. Isn't this a result affected by the how coarse the data is? I believe mobility means that the user appears in different zipcodes. Of course, 84% of sessions do not span more than 2 zipcodes.

- I believe that the conclusion of Section 3.1.4 is wrong. The data shows that the 50% of trajectories are AB, ABC, ABCD, ABCDE, etc... This does not imply that users are equally likely to end the day accessing the network from a location different than where they started as by returning back from the same location. For example, the data doesn't exclude the following possibility: an additional 25% of trajectories are of the form ABCB, ABCDB, ABCDC, ABCDEB, etc... This means that in 75% of trajectories, users don't end their days where they started.

- Section 3.2.1. I didn't know what "lower-bound" means.

- Figure 5. It's important to know how many users are found in each category. I assume that *very few* users see 50 base-stations, whereas more than half the users see 1 base-station only. This might skew the data and the findings.

- Is there any intuition why social networking apps dominate the x=10 point? I'm very skeptical about this finding.

- The \delta at the end of Section 3.2.1 confused the heck out of me. How is the analysis performed with a varying \delta? Does it mean that the analysis is repeated for different values of \delta, and the results agree? What's going on?

- The paper picks 3 as the magic threshold for how many locations are in a "comfort zone". But Figure 6 shows a very skewed preference for the top locations. Why not pick 1 as the threshold? It seems that the second location is almost 50% less popular than the 1st.

- I couldn't understand the intuition behind finding that the dating application is accessed from the top 3 locations, but not from home or work location. What's the 3rd comfort location?

- The methodology used in Section 4 to label hotspots as "day", "noon" does not preclude one hotspot from carrying two labels. How many such hotspots there are or are the hotspot sets completely disjoint? How do they influence the results? Is it a set of very popular hotspots that appear as "day", "noon", "evening", "night" and their workloads change based on time-of-day?

- I had a very hard time buying any of the results based on "encounters" given the coarse granularity of data. I don't encounter most of the people who are located in the same zipcode as I am or that share the same 3G access point.

- Section 6 claims that this paper is similar to the Bluetooth encounter studies because they study "the probability of people to meet each other". This is misleading -- Bluetooth's range is on the order of 30 feet whereas 3G's range is on the order of miles. It's very different.

==========

SIGCOMM 2009 Review #52D
Updated Monday 13 Apr 2009 1:17:59am EDT

Paper #52: The Genesis of Location-Aware Mobile Social Networking

---

Overall score: 3. Accept if room. Unsure about the overall rank of the paper. Need to compare with other borderline papers.
Reviewer confidence: 4. High confidence
Technical merit: 3. Average
Novelty: 4. New contribution
Relevance/community interest: 4. High
Longevity: 3. Not important now, possibly short lifetime in the future

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

This papers looks at a usage trace of mobile applications for 280K clients in a mobile network. The authors try to infer various properties from the trace using rule mining such as what applications are used in what locations. They also study generic mobility properties of users in the trace.

The trace is collected using cell towers so I assume all the data is 3G unless the capturing is being done on the client somehow.

===== Comments for authors =====

The title of the paper is misleading as I don't really see the connection with social networking. For that kind of a study you need knowledge of application level associations between clients and then measuring their interaction. This paper would be served better by being promoted as a study of mobile application usage characteristics as well as mobility characteristics of a population.

The location data used is very coarse grained. Given that most handsets still do not have GPS functionality which would be ideal for this study, even a triangulated position from something like Skyhook or A-GPS using cell towers would have significantly improved the inferences and value from the paper. Being associated to a celltower could give a very large radius and does not allow the understanding of the context in which an application may be used. If there is a coarse position, one may be able to associate the use of yelp type apps with dinnertime and being in a part of town with nightlife and restaurants. Most application usage has much finer grained context than what can be provided by cell tower association. If the cities were to deploy a bunch of femtocells it may be possible to better do some of these kinds of studies.

Some of these questions may require indoor or urban canyon location at a scale (<10m) which we may not get even with GPS, A-GPS skyhook etc (e.g. people looking up applications based on specific part of a mall that they are in). SO I understand that this trace may be the best we can do for now. However, the authors should acknowledge this and try to extract characteristics that are not unrealistic.
It is not clear how mobile social networking will make use of the conclusions in the paper. The location data is too coarse grained to figure out if local communication (ad hoc wifi, bluetooth etc) could be utilized for users with similar interests since they may be associated with the same cell tower but miles apart. These are the kinds of answers we need for modeling contact networking which can only be obtained with GPS tagged data.

The authors could potentially study how often users with similar interests (as gleaned from their data accesses) end up in the same cell tower. Such data is very useful for network operators who are deploying protocols to use broadcast content delivery techniques for popular content. This is an example of the kinds of questions the authors could try to answer that exploit their useful data set without reaching too far into conjecture.

==========

SIGCOMM 2009 Review #52E
Updated Wednesday 15 Apr 2009 2:47:27am EDT

Paper #52: The Genesis of Location-Aware Mobile Social Networking

---

Overall score: 3. Accept if room. Unsure about the overall rank of the paper. Need to compare with other borderline papers.
Reviewer confidence: 4. High confidence
Technical merit: 3. Average
Novelty: 3. Incremental improvement

Relevance/community interest: 4. High
Longevity: 3. Not important now, possibly short lifetime in the future

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

The paper studies traces from a mobile provider to understand if and how the application usage differs by location.

===== Comments for authors =====

This reviewer likes the goal of the paper to understand if the application usage of mobile users differs by location, time of day, etc....

However there are a few problems with the current paper and its approach:

First of all while this reviewer understands the data sensitivity very well it is somewhat confusing that you do not even roughly state when the trace was taken nor what kind of devices were monitored.
Ín addition it is confusing that you spend quite a bit of time explaining the Radius part of the data but do not mention what kind of data is underlying the application usage analysis?
Indeed, to this reviewer it is not obvious why its sufficient to only consider HTTP traffic. Is this appropriate?
It seems most email applications do not work on top of HTTP.
Moreover, on page 3 you talk about "we demonstrate that our dataset, despite its sampled nature, is capable of
accurately revealing human movement properties"
This is the first and only time that you mention sampling. What data do you sample in which manner?

Regarding the analysis it is confusing to this reviewer that you attempt to classify the overall mobility properties with data from just a single city. This implies that you do not have samples of movements across
longer distances. As such why is this sample representative? Why is the average meaningfull?

To judge the relevance of the later pieces of the analysis it would be good to understand how many users
fall in each category. Without this information its hard to judge the statistical significance.

==========

Comment
Paper #52: The Genesis of Location-Aware Mobile Social Networking

---

TPC meeting discussion:

Exciting paper, with a very unique dataset. Story is misleading though, and
authors don't quite deliver. The dataset has limitations due to coarse grained
mobility. Also, there aren't many details about the dataset: is it UMTS or
WiFi? How many hotspots/towers? How large is the area? It is also unclear what the

measurement errors are - a difficult problem though, as more context on the people / devices is needed.