**Nathan Eagle, PhD**

**Massachusetts Institute of Technology**
77 Massachusetts Avenue, Building E15-383
Cambridge, Massachusetts 02139–4307

**Reality Mining**
**http://reality.media.mit.edu**

Mobile +1 505 204 6637
Office +1 505 946 2768
Email nathan@mit.edu
http://web.media.mit.edu/~nathan

# The Reality Mining Data

## README.

The Reality Mining project was conducted from 2004-2005 at the MIT Media Laboratory. The Reality Mining study followed ninety-four subjects using mobile phones pre-installed with several pieces of software that recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status. Subjects were observed using these measurements over the course of nine months and included students and faculty from two programs within a major research institution. We also collected self-report relational data from each individual, where subjects were asked about their proximity to, and friendship with, others.

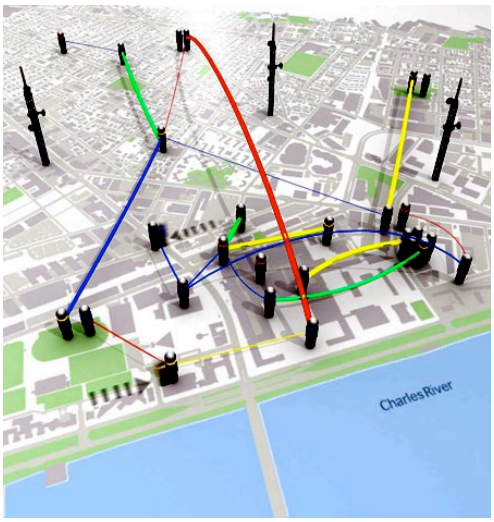


**Fig 1.** A visualization of the some of the Reality Mining data.

### Citation

If the data is used in a publication, please cite the following paper:

Nathan Eagle, Alex Pentland, and David Lazer. Inferring Social Network Structure using Mobile Phone Data, *Proceedings of the National Academy of Sciences (PNAS)*, 2009, Vol 106 (36), pp. 15274-15278.

### Subject pool

The subjects from this study consisted of students and staff at a major university during the months between September 2004 and June 2005. For this paper's analyses, we used a subset of the data collected for the Reality Mining study, incorporating the 94 subjects that had completed the survey conducted in January 2005 . Of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 subjects were incoming students at the university's business school. The subjects volunteered to become part of the experiment in exchange for the use of a high-end smartphone for the duration of the study.

### Mobile Phone Logging Software

The data for this paper came from Nokia 6600 phones programmed to automatically run the ContextLog application as a background process at all times. This application continuously logs passive behavior such as location (from cell tower ids) and other proximate subjects (from Bluetooth device discovery scans at five-minute intervals). The application also logs all of the phone's activity, including voice calls and text messages, active applications (such as the calendar or games), and the phone's charging status.

Data were collected from the phones using two methods. Approximately 30 of the subjects were provided data plans (GPRS) on their mobile phone. For this group we had the phones directly connect to our data server during the night and upload the new data logged during previous the day.

For the remaining subjects in the study, data was stored on each phone's internal 32MB memory card. The cards can store approximately four months of behavioral data before they need to be collected by the researchers.

An anonymized version of this dataset is currently available for download:

http://reality.media.mit.edu/download.php

## Data Description

*Phone log*

(TIME) 20060720T211505 (DESCRIPTION) Voice call (DIRECTION) Outgoing (DURATION seconds) 23 (NUMBER) 6175559821

*Bluetooth*

(TIME) 20060721T111222 devices: 000e6d2a3564 [Amy's Phone] 000e6d2b06ea [Jon's PalmPilot]

*Location*

(TIME) 20060721T111222 (CELL AREA) 24127, (CELL TOWER) 111, (SERVICE PROVIDER) AT&T Wirel (USER DEFINED LOCATION NAME) My Office

## Observational Accuracy

While the custom logging application on the phone crashes occasionally (approximately once every week), due to automatic restarts these crashes do not result in significant data loss. However, while the logging application can be assumed to be running anytime the phone is on, the dataset generated is certainly not without noise. Because we know when each subject began the study, as well as the dates that have been logged, we know exactly when we are missing data. These missing data are due to two main errors: data corruption and powered-off devices. On average we have logs accounting for approximately 85.3% of the time that the phones have been deployed.

## Inferring Location from Cellular Towers

A mobile phone has reception when it is within the range of a fixed cellular tower. While most cellular towers have ranges extending several square kilometers, in typical urban settings tower densities are significantly higher. Each tower has been assigned an ID that is logged by the mobile phones in our study. Using the tower IDs and respective transition timings (timestamps when the phone is handed off between cellular towers), it has been shown that a phone's position can be localized to within 100-200m in urban areas.

## Inferring Proximity from Repeated Bluetooth Scans

Bluetooth is becoming an increasingly popular short-range RF protocol used as a cable replacement to wirelessly connect proximate mobile electronic devices (such as phones and laptops) together. A key feature of a Bluetooth device is the ability to scan for other nearby Bluetooth devices. When a Bluetooth device conducts a discovery scan, other Bluetooth devices within a range of 5-10m respond with their user defined name (e.g.: Mark's 6680), the device type (Nokia Mobile Phone), and a unique 12-digit MAC hardware address (e.g.: 0012d186e409). A device's MAC address is fixed and can be used to differentiate one subject's phone from another, irrespective of the device name and type. When a subject's MAC address is discovered by a periodic Bluetooth scan performed by another subject, it is indicative of the fact that the two subjects' phones are within 5-10 meters of each other.

## Human Subjects Approval

Continuously recording a subject's daily behavior over an extended period of time has significant privacy implications. For example, under some circumstances, these data might be as sensitive as medical information. For IRB approval, we provided each subject with detailed information about

the type of information that would be captured and instructions how to temporarily disable the logging application. We also had strict protocols limiting access to the data. All personal data such as phone numbers were one-way hashed (MD5), generating unique ids used in the analysis. While we found that subjects were initially concerned about the privacy implications, less than 5% of the subjects ever disabled the logging software throughout the 9-month study.

### Constructing the Dyadic Observational Variables

Conducting periodic Bluetooth scans at 5 minute intervals generated approximately 4 million proximity events in the dataset. For each proximity event we have logged the two proximate MAC addresses, the current associated cellular tower for each of the phones, and the time and date of the event. The dyadic variables below come from these proximity events, as well as phone communication logs and the report survey data.

Because all of the phones are scanning every five minutes, if two subjects were together for 100 minutes there would be a total of 40 recorded proximity events. We therefore approximate each proximity event to be representative of a 2.5 minute time interval. To estimate the amount of proximity at a particular location such as 'Work', we multiply this time interval by the number of proximity events that involved the cellular towers associated with that location. A 'Proximity at Work' value of '15.7' for a particular pair of individuals would thus mean that during the times when their phones have logged the cellular towers associated with campus, the individuals have had an average estimated daily proximity of 15.7 minutes.

Data logged for each voice conversation on the mobile phone during the study included the time the conversation started, the duration and direction (incoming or outgoing) of the call, and the other phone number involved. If this other number was associated with another subject in the study, we incorporate the duration of the call into a statistic that estimates the average number of minutes of daily phone communication between each pair of subjects.

### MATLAB Network Survey Data

At the midterm of the 9-month study we conducted an online survey, which was completed by 94 of the 106 Reality Mining subjects. This survey included dyadic questions regarding the average reported proximity and friendship with the other subjects, as well as questions concerning the individual's general satisfaction with his or her work group. The questions used for this analysis are written below.

*Dyadic Questions*

• Estimate Your Average Proximity (within 10 feet) with Each Person at work / outside lab.
5 - at least 4-8 hours per day... 4 -at least 2-4 hours per day... 3 - at least 2 hrs - 30 minutes per day .... 2 - at least 10 - 30 minutes per day... 1 - at least 5 minutes .. 0 – 0-5 minutes (default)

***These data are represented in the network.lab and network.outlab matrices.***

• Is this Person a Part of Your Close Circle of Friends?
Yes / No (default)

***This data is represented in the network.friends matrix.***

*Note: The networks involve 94 subjects, however the data below involves 106 subjects. The indices in the networks (i = 1-94) are mapped to the subjects numbers (n = 1-106) subject subjects using network.sub_sort: network.sub_sort(i) = n. For example, network.sub_sort(2) = 4. That means the responses of subject 4 (s(4)) are shown in the 2nd row in the networks.*

### MATLAB Subject Data

The subject data involves 106 individuals, several of whom did not participate for a significant amount of time.

### *s(n).surveydata*

1. Have you travelled recently?
   1 Very often - more than a week/month 2 Often - week/month 3 Sometimes - several days/month 4 Rarely - several days/term 5 Never

2. Do you own a car?
   1 Yes 2 No

3. How many miles to you live from MIT?
   1. less than 1 2. 1-3 3. 4-10 4. more than 10

4. How do you daily commute to MIT?
   1. By foot 2. By bike 3. By T/bus 4. By car

5. How much has your social network evolved since the start of Fall term?
   1. A lot 2. Somewhat 3. Slightly 4. None

6. Have you been sick recently?
   1. Yes, in the last week 2. Yes, in the last two weeks 3. Yes, in the last month 4. No

7. How long into the term did it take for your social circle to become what it is today?
   1. Still evolving 2. 2 months into term 3. 1 month into term 4. Several weeks into term 5. First couple of days here

8. I use my phone:
   1. exclusively for work/school related matters 2. primarily for work/school related matters, but occasionally for personal/social use 3. equally for work/school and for personal/social use 4. primarily for personal/social use 5. exclusively for personal/social use

9. How often do you send text messages?
   1. Several times / day 2. once / day 3. once / week 4. once / month 5. never

10. The majority of my daily work communication is done through: (you can select more than one) face-face discussion
    1. Yes NaN. No

11. The majority of my daily work communication is done through: (you can select more than one) email
    2. Yes NaN. No

12. The majority of my daily work communication is done through: (you can select more than one) phone
    3. Yes NaN. No

13. The majority of my daily work communication is done through: (you can select more than one) text-messaging
    4. Yes NaN. No

14. The majority of my daily personal communication is done through: (you can select more than one) face-face discussion
    1. Yes NaN. No

15. The majority of my daily personal communication is done through: (you can select more than one) email
    2. Yes NaN. No

16. The majority of my daily personal communication is done through: (you can select more than one) phone
    3. Yes NaN. No

17. The majority of my daily personal communication is done through: (you can select more than one) text-messaging
    4. Yes NaN. No

18. I am satisfied with my experience at MIT thus far I am satisfied with my current social circle 1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

19. I am satisfied with my current social circle
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

20. I feel I have learned a lot this semester
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

21. I am satisfied with the content and direction of my classes and research this semester
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

22. I am satisfied with the support I received from my circle of friends
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

23. I am satisfied with the level of support I have received from the other members in my Media Lab research group / Sloan core team.
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

24. I am satisfied with the quality of our group meetings
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

25. I am satisfied with how my research group interacts on a personal level
    1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

### *s(n).mac*
The Bluetooth MAC address (unique hardware address) of the subject's phone.
'000e6d2a357b'

### *s(n).my_startdate*
The date the subject enrolled in the study.
'8/1/2004'

### *s(n).my_affil*
The subject's affiliation:
'mlgrad' – Media Lab Graduate Student (not a first year)
'1styeargrad' – Media Lab First Year Graduate Student
'mlfrosh' – Media Lab First Year Undergraduate Student
'mlstaff' – Media Lab Staff
'mlurop' – Media Lab Undergraduate
'professor' – Media Lab Professor
'sloan' – Sloan Business School

### *s(n).my_group*
The subject's research group.
'pattie'

### *s(n).my_imei*
The IMEI of the subject's phone:
[353383002009713]

### *s(n).my_neighborhood*
The subject's neighborhood.
'Porter'

### *s(n).my_hours*
The subject's reported hours at work.
'11am-8pm'

### *s(n).my_regular*
Does the subject report having a regular working schedule.
'somewhat'

### *s(n).my_hangouts*
The subject's reported hangouts
'restaurant/bar; friends'

### *s(n).my_predictable*
Does the subject report having a predictable schedule.
'very'

### *s(n).my_forget*
Does the subject report forgetting his phone at home / work?
'rarely'

### *s(n).my_battery*
How often does the subject report her battery runs out on the phones?
'occasionally'

### *s(n).my_sick*
How often does the subject report illnesss?

'rarely (once a year or less)'

### *s(n).my_sickrecently*
Has the subject reporting being sick recently?
'Yes, in the last week'

### *s(n).my_travel*
Does the subject report often traveling?
'Rarely - several days/term'

### *s(n).my_data*
The subject's data plan
'Unlimited

### *s(n).my_plan*
The subject's mobile phone plan
'national'

### *s(n).my_provider*
The subject's mobile phone provider
'AT&T'

### *s(n).my_minutes*
The number of minutes the subject buys each month.
'500'

### *s(n).my_texts*
How often the subject reports send text messages.
'rarely'

### *s(n).my_intros*
Whether the subject would like to receive introductions to others.
'often'

### *s(n).my_community*
How connected does the subject feel with her community?
'a little close'

### *s(n).comm*
Struct array with fields for each communication event. (Note that calls to the subject's own phone number is typically associated with checking voicemail.)

date: 732162.65994213 --Convert using datestr
event: 299 --Unique event ID
contact: -1 --The contact ID in phone's address book? (-1 = Not in address book)
description: 'Voice call' --Type of communication
direction: 'Outgoing' --Direction (Outgoing / Incoming)
duration: 0 --Duration in seconds (0 = didn't pick up)
hashNum: 165 --The hashed phone number of the other party

### *s(n).charge*
Date and time the phone is charging (1) or unplugged (0). (convert using datestr)

| | |
|---|---|
| 732339.674236111 | 0 |
| 732339.689803241 | 1 |
| 732339.696423611 | 0 |

### *s(n).active*
Date and time the phone has been in use (1) and not in use (0)

| | |
|---|---|
| 732338.463912037 | 0 |
| 732338.464340278 | 1 |
| 732338.465046296 | 0 |

### *s(n).logtimes*
Times when the logs were being written (not particularly useful)

732262.392453704
732262.392476852
732262.392476852

### *s(n).on*
When the phone is turned on (1) or off (0)

| | |
|---|---|
| 732322.839131944 | 1 |
| 732324.426944444 | 0 |
| 732324.442395833 | 1 |

### *s(n).locs*
Time-stamped tower transitions. [date, areaID.cellID] (0 is no signal)

| | |
|---|---|
| 732339.736053241 | 5188.40332 |
| 732339.737488426 | 5188.40811 |
| 732339.738287037 | 0 |

### *s(n).all_locs*
The unique set of towers seen by the subject. (areaID.cellID)

39402.30213
39402.30331
39402.30333

### *s(n).loc_ids*
An indexed version of s(n).locs. Towers are replaced by a unique ID.

1
842
842

### *s(n).device_names*
The names of the Bluetooth devices discovered on each scan.

'Solomon Biskers Computer'
'NORTHOLT'
'S25'
'MATTERHORN'
'S60'
'HOLUX GR-230'

### *s(n).device_macs*
The MAC addresses of the Bluetooth devices discovered on each scan. (Converted to ints using hex2num)

45452944210
58125624664
61960946218
10510993893
61965019994
34813389204

### *s(n).device_date*
The time / dates of each scan

732339.743506944
732339.747384259
732339.75125

### *s(n).device_list_names*
A list of all the devices names seen by the phone.
s(4).device_list_names{744} = 'HTHSV3a0189'

### *s(n).device_list_macs*
A list of all the devices Bluetooth MAC addresses seen by the phone (converted from hex to int)
s(4).device_list_macs{744} = 35197308840062

### *s(n).device_types*
The discovered Bluetooth device type (as determined by the standard Bluetooth protocol).

| | | |
|---|---|---|
| 1 | 3 | 129 |
| 1 | 1 | 16 |
| 2 | 1 | 640 |
| 1 | 1 | 400 |
| 2 | 1 | 640 |
| 31 | 0 | 0 |
| 1 | 3 | 912 |

### *s(n).device_list_types*
A list of the device types discovered by the phone

| | | |
|---|---|---|
| 2 | 3 | 640 |

2 3 640
4 1 256

### *s(n).cellnames*
An array of areaID.cellID and the string the user named the location.
[ 5188.48541] 'T-MobileLogan'
[ 5188.60291] 'T-MobileSwisshouse'
[ 5187.41803] 'T-MobileAmy'

### *s(n).apps*
The time each application was started and the total number of times the app was used.
all: {1x11060 cell}
snake_date: []
phone_date: [4430x1 double]
browser_date: [38x1 double]
camera_date: [92x1 double]
gallery_date: [73x1 double]
logs_date: [294x1 double]
clock_date: [307x1 double]
calendar_date: [6x1 double]
video_date: [7x1 double]
player_date: [5x1 double]
snake: 0
phone: 4430
browser: 38
camera: 92
gallery: 73
logs: 294
clock: 307
calendar: 12
video: 7
player: 5

### *s(n).timeon*
The total amount of time the phone has spent recording data (in days)
128.85751157417

### *s(n).app_dates*
The set of times when a user started an application
732339.738935185
732339.739027778
732339.753472222

### *s(n).home_ids*
The areaID.cellID of the tower we associate with the subject's home.
5123.40763 5188.40763

### *s(n).home_nights*
The nights when we find the subject at home.
732336.208333333
732337.208333333
732338.208333333

### *s(n).comm_local*
The total amount of local (Boston-based) communication events
558

### *s(n).data_mat*
Inferred locations at each hour of the day. 1 – home, 2 – work, 3 – elsewhere, 0 – no signal, NaN – phone is off
3 12 am - elsewhere
1 1 am - home
1 2 am – home ...

### *s(n).my_enddate*
The last date in the dataset
732339.745208333

### *s(n).comm_sms*

Number of text messages send and received
299

### *s(n).comm_sms_date*
A list of dates when a SMS was sent or received
732338.146956019
732338.880358796
732338.969236111

### *s(n).comm_voice*
Number of voice calls made and received
920

### *s(n).comm_voice_date*
The dates of the voice calls.
732340.042893518
732340.175914352
732340.176041667

### *s(n).comm_data*
The number of data sessions initiated on the phone
1570

### *s(n).comm_data_date*
The times when the data sessions were started
732339.85900463
732339.979710648
732340.051423611

### *s(n).places*
The distribution of times the subject was at home, elsewhere, work and with no signal.
home: [24x180 double]
elsewhere: [24x180 double]
work: [24x180 double]
nosig: [24x180 double]
all: [24x180 double]
startdate: 732160
endate: 732339.753506944
hours: [4315x1 double]
dow: [4315x1 double]
cell_vec: {1x4315 cell}
places_data: [4315x1 double]
off: [1220x1 double]
starton: [42x1 double]
endon: [42x1 double]

### *s(n).survey_start_n*
The date the subject started the survey.
10-Jan-2005

### *s(n).my_hashedNumber*
The subject's hashed phone number.
4

## Cellular Towers

*We do not have the actual locations of any cellular towers*. However, we do have the names each subject labeled the tower. From this, we can infer which towers are associated with 'Work' (MIT). These towers are the following:

> 5119, 40811, T-Mobile Media lab 1
> 5119, 40332, TMO Tech sq 2
> 5123, 40763, TMO MIT / Ashdown 3
> 5119, 40342, TMO Ashdown 4
> 5119, 40801, T-Mobile East campus / hyatt 5

> 5119, 40342, T-Mobile Inf corr 6
> 5119, 40802, T-Mobile Tang 7
> 5131, 43861, T-Mobile Tang 8
> 5119, 40793, T-Mobile Mit 9
> 24127, 132, AT&T Wirel 1-115
> 24127, 131, AT&T Wirel 1-115
> 24127, 2421, AT&T Wirel 2-103/ ML / End Inf cor
> 24127, 2353, AT&T Wirel Build 3
> 24127, 2833, AT&T Wirel Student center
> 24127, 111, AT&T Wirel ML / Mass Ave/ Infinite
> 24127, 182, AT&T Wirel Mass ave bridge 310 smoots / New house
> 24127, 2832, AT&T Wirel ML
> 24127, 113, AT&T Wirel Ml
> 24127, 2422, AT&T Wirel Ml
> 24127, 2833, AT&T Wirel Ml
> 24127, 112, AT&T Wirel Ml
> 24127, 2413, AT&T Wirel Ml
> 24127, 133, AT&T Wirel Ml
> 24127, 2433, AT&T Wirel Ml
> 24123, 261, AT&T Wirel Ml
> 24127, 2832, AT&T Wirel Medical
> 24127, 182, AT&T Wirel Mass ave bridge 310 smoots

### Date Discrepancies

You will see time-stamps that are Jan 1 2004, *ignore these*. This is what happens when the phone completely runs out of battery and needs to be reset. Use s(n).my_startdate to find when the subject joined the study – not the earliest date in the log file.