Submitted to *Working Manuscript*
manuscript (Please, provide the mansucript number!)

WORKING MANUSCRIPT. LAST UPDATE 08/01/16

The SmartBay project: connected mobility in the San Francisco Bay Area

Andrew A. Campbell
University of California, Berkeley, ,

Author2
Author2 affiliation, ,

The novel mobility-as-a-service paradigm, enabled by ICT and mobile computing, is changing the transportation landscape quicker than traditional data sources, such as travel surveys, are able to reflect. This is particularly true in the San Francisco Bay Area as the influx of citizens and businesses to the city, the volatility of job markets, evolving demographics and internal migration further increase the variability of mobility patterns. It is more important than ever to measure, realistically model and forecast travel demand in near real-time. The Smart Bay project extends the state-of-the-art in activity-based simulations by incorporating the anonymized data stream from the cellular network infrastructure. The efficacy of cellulary data in activity-based simulations is evaluated by comparing three scenarios: a control scenario using a traditional HHTS-based activity model, an activity model based only on cellular data, and a hybrid model using both the HHTS and cellular models.

Key words:
History:

1. Introduction

Informed transportation policy, planning, and operations decision making requires accurate forecasting of travel demand. Travel demand models are traditionally estimated using either stated or revealed preference travel survey data. The revealed preference survey is preferred, conducted in the form of an individual or household travel diary. These data collection methods provide a rich set of features for survey participants, but are limited in four critical ways. 1) One-off travel surveys are expensive, both in terms of monetary and time costs. 2) Sample sizes are very small compared to the populations they are supposed to represent. 3) Travel surveys are not longitudinal. In most instances, a travel diary is collected for only one or two weeks, making it difficult to capture seasonal effects. 4) Travel surveys are conducted infrequently. For example, the most recent household travel survey of the nine counties of the San Francisco metropolitan region was conducted in 2000 (BATS, the Bay Area Travel Survey) (MTC - Data Portal 2000). This data set still serves as the modeling baseline for many county and city agencies (though the data are augmented with even smaller project-specific studies) (MTC 2012). Given the accelerating pace of global urbanization, metropolitan economies and land-use evolve too rapidly to use traditional travel survey methods for building demand models. Transportation practitioners need more temporally relevant data sources.

The rise of ubiquitous mobile computing has created an opportunity to use large-scale passively collected location data that addresses these four shortcomings. Signals such as geotagged social media posts and cellular network data, provide a continuous, longitudinal stream of travel behavior information at a very large sampling scale. In particular, mobile phone data are a very attractive source due to both the scale of adoption and frequency of usage. In the United States, it is estimated that adult cell phone adoption rate exceeds 90% (Pew Research Ceneter 2015). Global cell phone adoption is estimated #TODO: TO BE SOME NUMBER I NEED TO FIND. Cell network providers collect daily user location data, providing a near real-time data stream. Since the data are linked with a unique account identifier, this stream becomes a longitudinal sample of mobility data for the lifetime of the cellular users account. Large volumes of cellular data records (CDRs) are generated daily and can be queried with relatively little time latency (as compared to travel surveys). This allows for the training and calibration of travel demand models using very large-sample, longitudinal data that is temporally accurate.

We sought to quantitatively evaluate the efficacy of using cellular data to train and calibrate travel demand models in the context of the San Francisco Bay Area. The metropolitan region consists of nine counties and 7.2 million inhabitants. An experimental approach was taken. As a baseline for model performance comparison, we used the Metropolitan Transportation Commissions (MTC) Travel Model to generate activity chains for a synthetic population. We used MATSim to simulate mobility and generate volume and link performance estimates for model evaluation. The baseline MTC Travel Model results were compared with two models using cellular data. One model used only cellular data to generate activity chains. The other was a hybrid model that combined the MTC Travel Model, American Community Survey census data, and cellular data.

We find that incorporating cellular data...

2. Background

Several studies have employed cellular data for analysis of human mobility (Wang et al. 2011) (Cho et al. 2011) (Frias-Martinez et al. 2012). These papers have focused on measures of travel distance (e.g. radius of gyration), or prediction of location sequences. However, the research into applying cellular data to transportation network analysis is very limited. A recent submission to MIT's NetMob conference Data for Development challenge explores using cellular data for estimating highway traffic volumes in Senegal (Liang and Frias-Martinez 2015). This work employs linear and support vector regression for prediction. The work we present is unique in employing microsimulation techniques for using cellular data for traffic analysis.

3. Data

The cases study encompasses the San Francisco Bay Area: a 7,000 sq-mi region spanning the nine counties under the jurisdiction of the Association of Bay Area Governments. The 2014 population of the Bay Area is estimated to be 7.5M residents (California Department of Finance 2016).

3.1. Network

The MATSim road network was created using OpenStreetMap (OSM) road network data, downloaded in July, 2015 (OpenStreetMap 2015). The user-generated OSM data offers very complete coverage in major metropolitan regions as well as rich feature sets including: link distance, number of lanes, speed limit, and hierarchical road classification. A manual inspection of dozens of freeway links in California found the OSM features to be accurate.

The data was clipped and filtered using Osmosis, an open source Java application for editing OSM data. The OpenStreetMap Standards and Conventions define tags for classifying roads hierarchically. There are 14 tags which encompass nearly all road links in the dataset. These range from motorway and trunk down to residential and smaller hierarchical classes (OpenStreetMap Wiki 2016). We found that for the Bay Area, the residential links constitute 74% of all links in the network. By filtering out the residential links, we were able to greatly improve the computational running time of MATSim without compromising regional-scale demand patterns. It is possible to maintain residential links for a localized area for future studies which require accurate neighborhood-level traffic reproduction. However, other limiting factors, such as the realism of MATSims queueing, traffic signal, and physics engines call into question the efficacy of including the lowest hierarchy links in the network

Once filtered, the geometry was simplified to a straight-line network to improve simulation speeds. Each intersection is a node, and a straight edge represents the road link connecting two intersections. To maintain realistic travel time skims, attributes of the original geometry network are preserved as attributes of link objects: length and free-flow travel speed. The final network used in the Smart Bay studies includes 564,368 links, 352,011 nodes and XXX miles of roads.

3.2. Demand

One of the essential input datasets for MATSim is a file describing the daily activity sequence for every agent in the population. For our experiments, we generated these demand files using three different methods. As a baseline, we used the Metropolitan Transportation Commissions (MTC) Travel Model. A second hybrid method combines MTCs Travel Model and the inferred activities from CDRs. The third method generates demand purely based on CDR activity inference. The data used in each method are discussed below.

3.3. MTC Travel Model

The MTC Travel Model is an activity-based demand model developed by Parsons Brinckerhoff, Inc. under contract for the MTC. It is a member of the Coordinated Travel - Regional Activity Modeling Program (CT-RAMP) family of models (Davidson et al. 2010). The model development, calibration and validation process is described in a 2012 report (MTC 2012). Agent populations were synthesized using historical and forecasted census and socio-economic distributions. The 2000 US Census Public Use Microdata Sample (PUMS) was used for generating empirical distributions of eight person types and four household types employed by the model. Aggregated TAZ-level socio-economic distributions from the year 2000 were provided by the Association of Bay Area Governments (ABAG). The baseline model used population distributions from the year 2000. Future scenario populations were drawn using IPF with forecasted distributions of TAZ-level person and household categories and socio-economic variables.

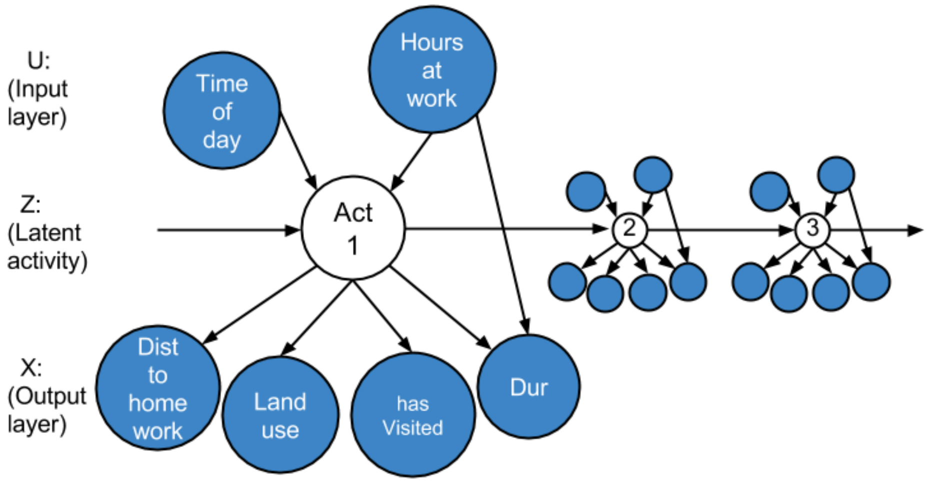
The activity segmentation was based on the 2000 Bay Area Travel Survey (BATS). The 16 original activity categories from BATS were aggregated into 10 types for the Travel Model. All major agent decision making, excepting traffic assignment were modeled using a sequence of multinomial logit choices ranging in scope from work and school location to intra-tour mode choices. MTC Travel Model was calibrated and validated against the Caltrans State Highway traffic count databases.

For our baseline modeling benchmark, we sample agent activity chains directly from the MTC Travel Model.

3.4. Hybrid Cellular-MTC Model Data

For the hybrid demand model, we used the MTC Travel Model (detailed above) and cellular data in the following ways:

- **Primary activities** Home and Work locations detected from cellular data were randomized within each TAZ and adjusted (scaled up) according to the total population estimates available from ABAG as published within the MTC Travel Model One.



7

Figure 1 **IO-HMM Specification (Yin et al. 2016)**

- **Secondary activities** While the developed methods allow us to generate secondary activities for individuals (see section 4.1 above), further research is required to guarantee that samples generated from the model will provide guarantees in protecting users locations privacy. Particularly, in case of the IO-HMM model that is trained on historical movement data of a single individual, it may over-fit a set of activities and locations in a way that in its entirety it will be unique to the given user (a set of locations and transitions may represent a mobility fingerprint for an individual). While this risk is minimal, in the hybrid model that we describe, a set of secondary activities was replaced by the fully synthetic tours generated with the MTC model. Further research is required to understands privacy guarantees and aggregation trade-offs in modelling detailed travel itineraries of travellers.

3.5. Pure Cellular Demand Model Data

In our third experiment design, we only use the generative IO-HMM model, trained on cellular data, for producing agent activity chains, Figure 1. A detailed discussion of this model is beyond the scope of this paper. Interested readers may refer to Yin et al. (2016). A brief description of the model and its application to activity chain generation follows.

3.6. Performance Data

Calibration and validation of the traffic volumes was conducted using freeway traffic counts from the Caltrans Performance Monitoring Systems (PeMS) (Caltrans 2015). We downloaded all 5-minute count data for Caltrans District 4 for the complete months of June, July and August 2015. Data for a total of 3,322 vehicle detector stations (VDS) was collected.

Although there as thousands of VDSs in District 4, there are many issues in the data quality to consider. During the study period, less than half of the stations were fully functional. Although the PeMS system employs heuristics to estimate missing data, we wanted to calibrate and validate against only against true data. To this end, we employed six levels of filtering:

1. **Date Range** - Only stations that were active during the study period were used.
2. **Link Matching** - Only sensors that could be matched to within 100 feet of a link in the MATSim network were kept.
3. **Missing Data** - Although a station may be active during the study period, many have significant periods of no reported counts. We removed all stations with more than one hour of missing data.
4. **Boundary Buffer** - To avoid edge effects, we removed any stations that were within 10-km of the study are boundary.
5. **Outlier Detection** - There are frequent "false outliers" in the PeMS data. These days are unusual due to errors in the PeMS reporting system instead of actual changes in traffic patterns. For example, a station may be reporting 100% of lanes are observed, but it will report the same volume of traffic in every time bin for that day. Another common error is gross under reporting of flow even when no accidents or construction were present. We used a one-class support vector machine (detailed below) to identify and remove stations with false outliers.
6. **Fraction Observed** - The 'Observed' attribute in the data describes the fraction of lanes for which the detectors were working during that reporting period. We only kept stations where at least 50% lanes were observed during every day of the study period.

After applying the six filters, only 774 usable stations remained. In Table 1 we see the results of sequentially applying the filters. It is important to note that each removed station is only attributed to one filter. Once a filter was triggered, the station was removed and the following filters were not applied. Although the total number of removed stations would stay the same, changing the order of filter applications would change the results in Table 1.

Table 1 **VDS Filtering Results**

Date Range	206
Link Matching	119
Missing Data	216
Boundary Buffer	185
Outlier Detection	200
Fraction Observed	1,622

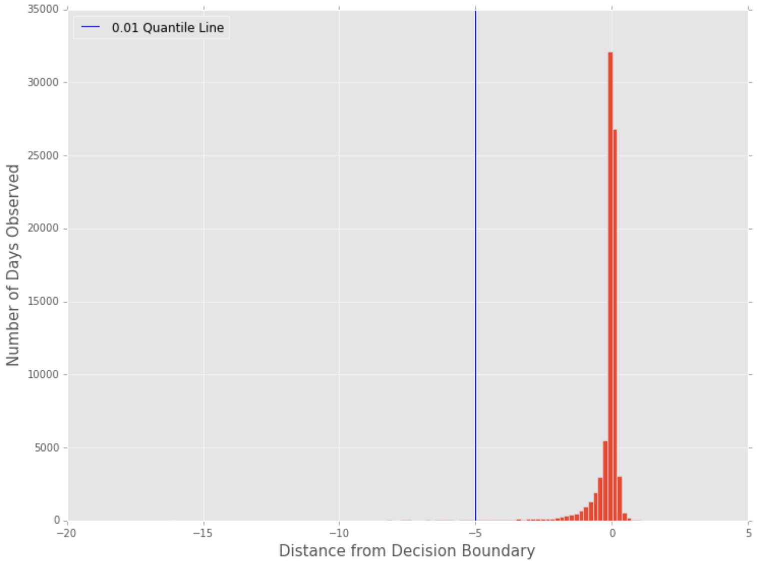


Figure 2 **One-Class SVM Outlier Detection - Distribution of Distances from Separation Plane**

3.6.1. One-Class SVM Outlier Detection A one-class SVM was used to identify "false outliers" among each station that satisfied the first four filters. For a single station, we created scaled and centered feature matrix, X. A single day of hourly volumes constitutes one point in a 24-dimension space. The feature matrix is thus in $\mathbb{R}^{n\times 24}$ space, where n is the number of days in the study period. We train the SVM on X to define a separation plane. The most typical days fall within the space defined by the plane and have a positive distance from the plane. The outliers fall outside and have negative distance. The more negative the distance from the separation plane is, the more unusual a day is. From Figure 2, we see that 99% of days are closer than -5.0 from the plane.

We manually inspected the data for 10 of the stations found to have days beyond the -5.0 distance. It was decided that if more than 5% of days for that station were beyond that distance, then the station would be removed.

4. Methodology

We evaluated the efficacy of using cellular data for modeling travel demand by running microsimulation experiments on the MATSim platform. The control scenario was a draw of 500,000 typical-weekday plans from the MTC Travel Model. We compared validation results with two treatments: the Hybrid Cellular-MTC Model (section 3.4) and the Pure Cellular Demand Model (section 3.5).

4.1. MATSim Platform

The MATSim (Multi-Agent Transport Simulation) platform is an agent-based activity model that performs microscopic modeling of traffic (using link performance functions) and agent decision

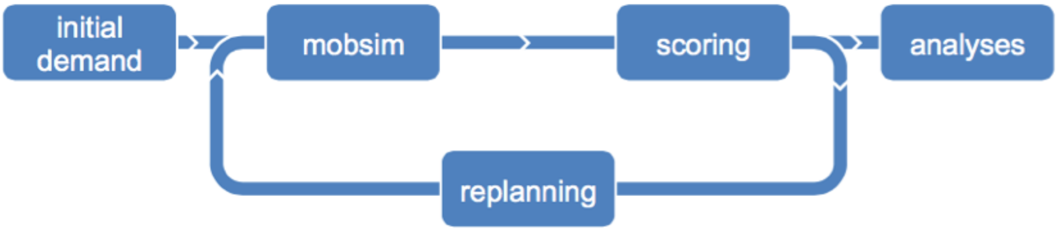


Figure 3 **The MATSim Cycle (Andreas Horni et al. 2016)**

making (Andreas Horni et al. 2016). The MATSim run cycle, Figure 3, is an iterative process whereby agents make adaptations to routing, activity timing, and other optional choices until convergence is reached. As input, each agent is assigned an activity chain (initial demand), complete with activity types, timing and location. During the mobility simulation (mobsim), the agents travel the network, interact, and experience congestion which lowers their overall utility scores for the day. During the replanning phase, a subset of agents may adapt their routes and activity timings. For our simulations, we restricted replanning adaptation to random selection of 10% of the population during each iteration. Many other forms of adaption are possible with MATSim, but for this project we have restricted adaptation to timing and routing. Agents incur a negative penalty for deviating from their original activity timings, so dramatic shifts in activity start and end times are not possible. Rerouting agents are allowed to update their routes to the new shortest path, based on the loaded network conditions in the most recent mobility simulation.

We used the hybrid MTC-CDR activity model to generate initial demand for a typical weekday. The scenarios simulate a single 24-hour day for 463,000 agents, scaled up to represent the total driving population. For this initial implementation, we cast all agent demand into private passenger-car equivalents. Thus each virtual car only carries a single agent, but it may represent more than one passenger trip.

4.2. Microsimulation Calibration

For calibration, we used the full set of 1,166 PeMS sensor stations that could be matched to the MATSim network. Simulation performance was evaluated using the root mean squared error (RMSE) calculated over all sensors over all hours of a 24-hour day:

$$RMSE = \sqrt{\frac{\sum_{s\in\mathcal{S}}\sum_{h=1}^{24}(\alpha\hat{y}_{s,h} - y_{s,h})^2}{\sum_{s\in\mathcal{S}}\sum_{h=1}^{24}1}} \tag{1}$$

where $\mathcal{S}$ is the set of all vehicle detector stations used in calibration, and $\hat{y}_{s,h}$ and $y_{s,h}$ are the simulated and measured counts for station s at hour h. The scalar coefficient α is a MATSim tuning parameter called the Counts Scale Factor. Since the number of agents may not be the same as the real population, the Counts Scale Factor is used scale simulated counts to best match the real world counts. The Counts Scale Factor has no impact on the actual microsimulation and is thus adjusted post-simulation to minimize the RMSE. Taking the derivative of equation 1, the optimal factor, α^*, is found:

$$\alpha^* = \frac{\sum_{s\in\mathcal{S}} \sum_{h=1}^{24} \hat{y}_{s,h} y_{s,h}}{\sum_{s\in\mathcal{S}} \sum_{h=1}^{24} \hat{y}_{s,h} \hat{y}_{s,h}} \tag{2}$$

Rahka et al. define model calibration as the selection of ”input parameter values that reflect the local study areas network, climactic, and driver characteristics” (Rakha et al. 1996). Driver characteristics calibration enters into the hybrid MTC-CDR demand generation described in previous sections. It also enters within MATSim. Agents evaluate their day of experienced activities and trips using the Charypar-Nagel scoring function; a utility function tailored for the co-evolutionary learning algorithm that the agents employ (Andreas Horni et al. 2016). The key behavior of this scoring function is a trade-off between the positive score accrued performing activities, and the negative score from traveling. We tuned several agent scoring parameters during calibration: how sensitive agents are to starting activities later than scheduled, the penalty for ending early, and the disutility of travel. These parameters were found to be of second order importance compared to changes in network performance.

MATSim provides two main levers for calibrating network performance: the flow capacity factor and the queue storage capacity factor. The flow factor dictates how rapidly link travel speed decays with volume. The link storage factor controls the link density constraints, which determine the acceptance rate of a link for incoming vehicles. The role of the two factors is succinctly described by Nurhan et al. (2003): ”link outflows are constrained both by the flow capacity of the link itself and by space limitations on the receiving link” (Cetin et al. 2003). For calibrating these factors, a good initial guess is to take use the agent-to-population ratio. So if we have a 10% sample, a good initial guess is 0.10. However, the complexity of the simulation prevents easy prediction of the impacts of adjustments to these factors and an iterative guess-and-check approach is required. For our population of 463,000 agents, we found that 0.12 worked
best for both scale factors.

The final calibration parameter is the counts scale factor. This parameter does not actually impact the agent behavior or link performance. It simply scales the simulated the counts up to match the observed volumes. After a simulation runs to convergence, we adjust the counts scale factor such that the total simulated and observed counts match. We chose a final value of 13.35, meaning each agent vehicle represents 13.35 observed vehicles.

4.3. Validation

Since we used the MTC Travel Model in our hybrid demand model, we sought to reproduce same validation metrics employed by that model's creators. We have attempted to reproduce the the measures in Tables 68 - 70 in the MTC Travel Model calibration and validation report (MTC 2012). These tables describe daily, AM peak, and PM peak predicted and observed flows at 27 key screenlines located at county borders and bridges. The AM peak is defined as 6:00 - 09:59, and the PM peak is 16:00 - 18:59. In Tables **??**, **??**, and **??**, we have used our PeMS typical weekday profiles for the observed values and the MATSim simulated volumes for the predicted counts. We have included 'NA' placeholders for locations that were included in the MTC report, but for which no PeMS data was available.

5. Results and Conclusions

The SmartBay platform serves two purposes. First, it is a test bed for forecasting real-world conditions and what if scenario testing for operational decision-making. For this purpose, the model is validated to meet the Federal Highway Administrations accuracy specifications. SmartBay also provides an experimental platform for exploring future mobility trends. We test scenarios where an on-demand service greatly reduces travel costs and find that social influence comes to dominate the choice of location for discretionary activities. We also explore shifts in modality driven both by direct social pressure and an indirect tendency to behave similarly to ones immediate peers.

Acknowledgments

References

Andreas Horni, Kai Nagel, Kay Axhausen, eds. 2016. *The Multi-Agent Transport Simulation: MATsim*. URL `http://www.matsim.org/docs/userguide`.

California Department of Finance. 2016. E-1 Population Estimates for Cities, Counties, and the State - January 1, 2015 and 2016. URL `http://www.dof.ca.gov/research/demographic/reports/estimates/e-1/view.php`.

Caltrans. 2015. Caltrans PeMS. URL `http://pems.dot.ca.gov/`.

Cetin, Nurhan, Adrian Burri, Kai Nagel. 2003. A large-scale agent-based traffic microsimulation based on queue model. *IN PROCEEDINGS OF SWISS TRANSPORT RESEARCH CONFERENCE (STRC), MONTE VERITA, CH*. 3–4272.

Cho, Eunjoon, Seth A Myers, Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1082–1090.

Davidson, William, Peter Vovsha, Joel Freedman, Richard Donnelly. 2010. CT-RAMP family of activity-based models. *Proceedings of the 33rd Australasian Transport Research Forum (ATRF)*. URL `http://atrf.info/papers/2010/2010_Davidson_Vovsha_Freedman_Donnelly.pdf`.

Frias-Martinez, Vanessa, Jesus Virseda-Jerez, Enrique Frias-Martinez. 2012. On the relation between socio-economic status and physical mobility. *Information Technology for Development* **18**(2) 91–106. doi: 10.1080/02681102.2011.630312. URL `http://dx.doi.org/10.1080/02681102.2011.630312`.

Liang, Tony, Vanessa Frias-Martinez. 2015. Cars and Calls: Using CDR Data to Approximate Official Traffic Counts. *NetMob D4D Challenge submissions*. URL `http://www.vanessafriasmartinez.org/uploads/d4dPaperFrias.pdf`.

MTC. 2012. Travel Model Development: Calibration and Validation. URL `http://mtcgis.mtc.ca.gov/foswiki/pub/Main/Documents/2012_05_18_RELEASE_DRAFT_Calibration_and_Validation.pdf`.

MTC - Data Portal. 2000. Bay Area Travel Survey (BATS). URL `http://dataportal.mtc.ca.gov/bay-area-travel-survey-bats.aspx`.

OpenStreetMap. 2015. Mapzen metro extracts - San Francisco Bay. URL `https://mapzen.com/data/metro-extracts/`.

OpenStreetMap Wiki. 2016. Highways. URL `http://wiki.openstreetmap.org/wiki/Highways`.

Pew Research Ceneter. 2015. Device Ownership Over Time. URL `http://www.pewinternet.org/data-trend/mobile/device-ownership/`.

Rakha, Hesham, Bruce Hellinga, Michel Van Aerde, William Perez. 1996. Systematic verification, validation and calibration of traffic simulation models. *75th Annual Meeting of the Transportation Research Board, Washington, DC*. Citeseer. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.1387&rep=rep1&type=pdf`.

Wang, Dashun, Dino Pedreschi, Chaoming Song, Fosca Giannotti, Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1100–1108.