Revisiting the Generality of the Rank-based Human Mobility Model

Darshan Santani
Idiap Research Institute
EPFL, Lausanne, Switzerland
dsantani@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPFL, Lausanne, Switzerland
gatica@idiap.ch

UbiComp '13 Adjunct, Sept 8-12, 2013, Zurich, Switzerland.
ACM 978-1-4503-2139-6/13/09...$15.00.

Abstract
Location-based social networks, in addition to revealing users' online social network, also informs users' actual movements in the offline physical world. Due to this, they have recently been used in large-scale mobility and urban studies. In this paper, using a rigorous statistical methodology, we have found that a rank-distance distribution, which in recent influential research has been suggested to be a "universal" mobility law across cultural, demographic and national boundaries, does not follow a power-law distribution as originally claimed. Using a large-scale dataset obtained from Foursquare in Switzerland and New York City, we have shown that place transitions can be better explained using a log-normal and power-law with exponential cutoff model. Our study suggests that urban mobility patterns are more nuanced than previously reported and that goodness-of-fit tests need to be done in view of the generality of human mobility models.

Author Keywords
Urban Mobility, Location-based Social Networks, Foursquare

ACM Classification Keywords
H.2.8 [Database Management]: Database Applications, Spatial databases and GIS, Measurement.

1 Introduction

In the last 5 years, sensor-rich mobile devices and online social networks have experienced an exponential rise. Fueled by their growth, location-based social network (LBSN in short) have emerged, which combine the sensing capabilities of smartphones with the friendship structure of online networks. LBSNs have opened new paths to understand linkages between the virtual and physical worlds. In addition to revealing information about users online social network, these networks also inform about their actual movements in the physical world. The combination of network structure along with highly granular mobility information provide unprecedented amount of user behavioral data to study and model human mobility patterns at a scale which was not feasible before.

In a typical LBSN, users voluntarily announce their arrival to a given place (a process known as check-in) and share information about their visits to places with everyone in their friendship circle. Given LBSNs current user base and popularity, it is fair to say that they are young and still evolving. Typically, LBSNs are manual check-in systems [6], wherein users, based on their personal preferences and various social/monetary incentives, have complete control over when, where and to whom they would like to disclose their place visitations.

In the research community, all empirical analyses on LBSNs so far have been based on a sample of global movements of people in space and time due to the inherent nature of data collection [8, 2, 6]. Due to restrictive programming interfaces, primarily due to privacy concerns, it is not possible to publicly collect user-level mobility data for all users of a location-based service. For instance, Foursquare, the world's leading LBSN, does not allow to collect user-level check-in data. Typically, a sample of Foursquare public check-ins are instead collected via Twitter's streaming API for all those users who post their check-in updates on their Twitter account as well [9, 2]. As a result, data obtained from services like Foursquare, represent only a fraction of human movement patterns. In this paper, we assume that mobility patterns captured via LBSNs are a representative sample of global movements and thus useful to study human urban mobility.

Studying human mobility has been an area of active research in recent times [1, 5, 12]. There is a substantial body of work – qualitative and quantitative – in the literature which have looked into how people move around in the physical world and the inherent motivation and costs associated with mobility [10, 13]. Analyzing individual mobility patterns – especially in urban areas – is crucial from a societal point of view as it has implications to transport planning, urban studies and management, epidemic spread and emergency response. As a result, a lot of effort in the research community has gone into finding empirical laws and models to characterize the heterogeneity of human movements across different urban regions. In this paper we take a critical look into one such model for human mobility.

A recent influential study has identified universal urban human mobility patterns using data obtained from Foursquare [8]. This study has proposed that the probability of visiting a place, when measured as a function of *rank-distance* (i.e., the number of intermediate places between source and destination, as opposed to mere physical distance between them), exhibit consistency which cuts across cultural and national boundaries. In this paper we statistically examine whether this particular rank-based model holds true or not on independently collected data. To undertake this analysis, we have collected and analyzed Foursquare dataset con-

sisting of more than 660,000 check-ins from within Switzerland and New York City.

In summary, we address the following research questions:

1. Does the *rank-distance* follow a power-law like distribution, as suggested in earlier research?
2. If it does not follow a power-law like distribution, which other heavy-tailed distributions can better describe transitions between places?

To answer these questions, we base our statistical analysis following the seminal work by Clauset et al. [3]. Their paper provides a statistical framework to estimate power-law fit for empirical distributions, compute the goodness-of-fit tests for a power-law like behavior, and statistically compare and evaluate alternate heavy-tailed distributions in favor of (or against) a power-law distribution.

2 Existing Models of Human Mobility

Modeling and analyzing human movement patterns have been an area of research and debate in the scientific community. Various models of human mobility have been proposed in the literature, ranging from distance-based models to gravity-based models to rank-based models. In recent times, due to the availability of large-scale datasets obtained from mobile sensing and online social networks, it has become possible to validate these models at a scale and a spatial resolution which was not feasible earlier. In this section we highlight two well-known models which have been proposed to capture the heterogeneities of human mobility.

2.1 Distance-based Model

The first model states that the probability of moving between places decreases as the geographical distance between the locations increases. In other words, the probability of traveling from source (s) to destination(d), $P[s \to d]$, decreases as a power of distance between them, $r(s,d)$. Mathematically it is given by:

$$P[s \to d] \propto r(s,d)^{-\alpha}$$

In recent times, one of the influential works to empirically validate this model has been reported in [5]. Using large-scale cellular data records (CDR) obtained from mobile operators, the authors propose that human *displacements* are well approximated by a truncated power-law distribution (a.k.a. power law with exponential cutoff, also see Section 4.3) with scaling exponent (α) equal to 1.75 (± 0.15).[1]

2.2 Rank-based Model

A second model, recently proposed by [8] states that absolute physical distance is not the decisive factor in modeling human displacements. Instead, they suggested a rank-based model inspired by Stouffer's theory of *intervening opportunities* [13], which says that the probability of traveling from source to destination is directly proportional to the number of opportunities closer to source than destination. Mobility thus is driven by a spatial distribution of opportunities, as opposed to mere physical distances. This model further proposes that transition probability varies inversely as a power of rank [8]. Formally, the rank of a transition is defined as the number of intermediate places which between source and destination. As per the rank-based model, for a scaling exponent α, the transition probability from source (s) to destination (d), $P[s \to d]$ is described as:

$$P[s \to d] \propto rank_s(d)^{-\alpha}$$

[1] In the literature, variants of the distance-based model have been proposed as well, but we are omitting their details due to lack of space.

where $rank_s(d)$ is defined as the total number of places geographically closer to source than the destination. Transitions with a place rank of 1 implies that user has checked in to the same place again, i.e., $rank_s(s) = 1$ for all places.[2]

As stated in the introduction (Section 1), Noulas et al. [8] have reported that the distance-based model for human mobility does not exhibit universal properties, but instead human transitions are better explained using a rank-based model. In other words, their paper has suggested that the rank-distance distribution follows a universal power-law model.

3 Dataset

In this paper, we present our analysis based on check-in dataset from Foursquare. Foursquare currently reports having over 3 billion check-ins from over 30 million users worldwide [4]. To respect users' privacy, Foursquare does not provide any direct mechanism to gather user-level check-in data. A common practice, therefore, is to collect check-ins via Twitter streaming API for all those users who post their check-in updates on their Twitter account as well. We have used this workaround to gather our dataset. Our current analysis is based on two different datasets, which are described below:

1. **Swiss Check-in Dataset**: We collected check-in data within Switzerland (CHE) using the data collection methods described above. The dataset spans more than 62,000 check-ins from 15,845 users over a period of 6 months between December 2011 and June 2012. In addition, we have also analyzed the movement trajectories from within the Zurich (ZRH) canton, which includes the largest Swiss city (Zurich) and its vicinity. Within Switzerland, ZRH canton has the largest Foursquare contribution amongst all 26 cantons, comprising of more than 30% of national check-ins.

2. **NYC Check-in Dataset**: The second dataset is obtained directly from Cheng et al.[2], which spans 22 million check-ins from 220,000 users across the globe. In this paper, we restrict our analysis to check-ins from New York City (NYC) only. The NYC dataset consists of over 600,000 check-ins over 318 days starting in March 2010.

Table 1 lists the basic statistics of datasets summarized above. Based on these statistics, it is easy to compare the relative popularity of Foursquare in New York City with Switzerland.

	ZRH	CHE	NYC
Number of Users	2,003	4,968	19,294
Number of Places	4,078	15,845	18,612
Number of Check-ins	19,333	62,714	602,898
Period of Analysis (days)	185	185	318
Area (in km^2) [15]	1,729	41,285	784

Table 1: Summary Statistics of Foursquare Dataset

4 Analysis

Now that we have described the Foursquare dataset in detail, in this section we present our rigorous statistical analysis. First, we fit a power-law model to the dataset, then we perform the goodness-of-fit tests to statistically validate

[2] To look at the results with $rank_s(s) = 0$, as defined in Noulas et al. [8], refer to the supplementary material here: http://idiap.ch/~dsantani/mobility/

the power-law hypothesis, and last but not least we evaluate alternate heavy-tailed distributions in favor or against the power-law distribution.

4.1 Fitting Power Law to Foursquare Data

In this section we focus our attention towards fitting the power-law distribution, in particular computing the scaling exponent α for our dataset.

4.1.1 Estimating the Scaling Exponent

We begin our analysis assuming a power-law like distribution for transition ranks. We compute ranks for every place transition and approximate a power-law fit using the methods described in [3]. More precisely, we approximate our discrete place-rank dataset to be a continuous distribution, and apply the method of maximum likelihood to estimate the scaling parameter, as given by the following equation (For mathematical derivations and proofs, the reader is referred to [3]):

$$\alpha \simeq 1 + n \left[\sum_{i=1}^{n} ln \frac{x_i}{x_{min} - \frac{1}{2}} \right]$$

In the above equation, x_{min} indicates the place rank x, where the power-law scaling begins. A priori we do not know the place rank where the scaling begins, and moreover we are interested in estimating the fit for the complete dataset. So, we have set x_{min} to be 1 for now. With this choice, we have obtained a scaling exponent $\alpha = 1.18, 1.18$ and 1.16 respectively for Switzerland, Zurich and NYC respectively, shown in Table 2. These values are similar to exponent values of 0.88 and 0.93 obtained in [8] and [6], which points towards the similarity of our dataset with ones used in these earlier studies.[3]

Note that in our analysis we have adopted the definition of a place rank as described in Section 2. Transitions with a place rank of 1 implies that user has checked in to the same place again. From Table 2, we observe that in NYC more than 25% of place ranks ($x_{0.25}$) are 1. That is, one quarter of place transitions in NYC are happening to the same venue, albeit at different times. While in CHE, consecutive visits to the same venue happen in over 10.5% of total transitions. We can think of one possible explanation for this phenomenon: Foursquare provides monetary and social incentives (badges, crowns, mayorship, etc.) to users who have performed the maximum number of check-ins to a given place. Due to the inherent game mechanics, users are incentivized to check in to the same venue time and again. This trend might simply be more popular in NYC than in Switzerland.

	x_{total}	$x_{0.25}$	$x_{0.50}$	α	p
ZRH	17,330	11	216	1.18	0.00
CHE	57,746	12	158	1.18	0.00
NYC	583,604	1	1,171	1.16	0.00

Table 2: Summary statistics for different regions. x stands for a place rank given a transition within the region. $x_{0.25}$ and $x_{0.50}$ represent the first and second quantile of place ranks respectively. α indicates the power-law exponent fitted to the entire dataset, with the given p-value. Statistically significant p-values are shown in **bold** (i.e., no statistical significance is found for the power-law fit)

4.1.2 Visual Inspection

A typical characteristic of power-law behavior is that if the underlying variable is distributed as per the power-law, then

[3] To the best of our knowledge, [8] and [6] have computed the scaling parameter on the whole dataset i.e., by setting x_{min} to 1.

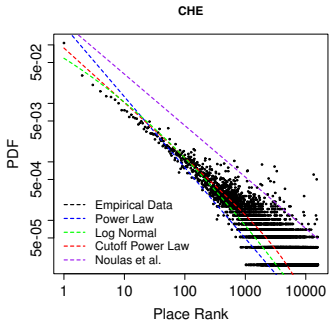

Figure 1: PDF of *rank-distance* on log-log scale for Switzerland

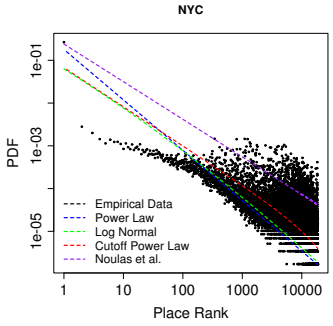

Figure 2: PDF of *rank-distance* on log-log scale for NYC

the probability distribution function (PDF) and the complementary cumulative distribution function (CCDF) will be a straight line on log-log axes. It is important to realize that having a straight line on a log-log PDF plot is a necessary but not sufficient condition for power-law like behavior [14].

In order to visually reveal the power-law nature of the rank-distance distribution, Figure 1 and 2 show the PDF of place ranks on logarithmic scale for Switzerland and New York City. Due to space constraints, we have omitted the PDF plot for Zurich, though the plot looks similar to the one obtained for Switzerland. Figure 1 and 2 also demonstrate fits for competing heavy-tailed distributions (discussed in Section 4.3). The blue curve indicates power-law fit to the data, with the green curve indicating a log-normal fit, while the red curve highlighting a power-law fit with exponential cutoff. We also show a purple curve which depicts the "universal" scaling exponent (with α as 0.88 and intercept value of 0.24) as estimated by [8][4]. Based on visual inspection alone, it appears that power-law is not the most suitable fit; rather power-law with exponential cutoff provides the best possible fit, though log-normal also looks like a better fit when compared against a pure power-law. (We will discuss this issue in detail in Section 4.3)

It is evident from Figures 1 and 2 that the PDF plot is noisy in its right fat tail, due to a sudden drop in the number of high-ranked transitions. More than 73% of all transitions in Switzerland happen to destinations with a rank of less than 1000, while it is over 48% for NYC. Due to the inherent noise in the tail, it is often useful to consider the CCDF of a power-law distributed variable. We show the CCDF plots in Figures 3 and 4 along with the respective distribution fits (as in Figures 1 and 2). Again from these figures, it is evident that the cutoff power-law model provides a better fit for the transition-rank data.

[4]Noulas et al. [8] has used "least squares based optimization" to estimate the scaling exponent

4.1.3 *Estimating the Lower Bound Parameter*

While estimating the scaling exponent in Section 4.1.1, we have assumed x_{min} to be known, and set its value to 1 in order to estimate the fit for the whole dataset. In this section, we relax this assumption and compute an optimal x_{min} where the power-law scaling begins, assuming a power-law like distribution for data above x_{min}.

As in Section 4.1.1, we have followed the statistical methods described in [3] to compute the lower bound for the scaling region. In brief, we choose x as x_{min}, which minimizes the distance between probability distributions of observed empirical data and the best-fit power-law model. The Kolmogorov-Smirnov (KS) statistic [7] is used to measure the distance between the respective distributions. In practice, we iterate through all possible values of x and compute KS statistic (denoted by D) between our data and the model that best fits the data above x_{min}. Once we have D values for every x, the x which minimizes D is our lower cut-off parameter x_{min}, and the exponent α corresponding to x_{min} is the power-law scaling exponent α in the region $x \geq x_{min}$.

Table 3 lists the respective x_{min} and α for all regions under investigation. We make several observations. First, it is evident that these regions have different power-law exponents, even when we account for different scaling regions. Second, we observe that for all regions, x_{min} is significantly large relative to the maximum possible rank (x_{max}). For NYC with $x_{min} = 8266$, we are only fitting the model to about 19% of the dataset; while for CHE, the fit is to only 15% with the computed lower bound. Given that the dataset is quite noisy in its right tail (Figures 1 and 2), it is hard to expect a pure power-law a possible explanation

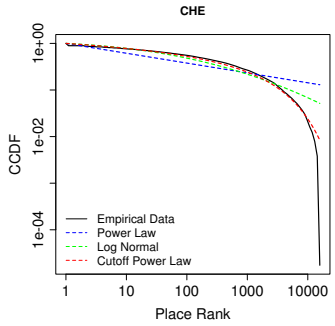

Figure 3: CCDF of *rank-distance* on log-log scale for Switzerland

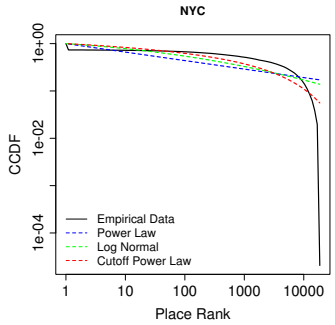

Figure 4: CCDF of *rank-distance* on log-log scale for NYC

to model place transitions on, even if it is able to explain, it is definitely not applicable to all place transitions.

	x_{total}	x_{max}	x_{min}	α	p
ZRH	17,330	4078	1,527	2.97	0.00
CHE	57,746	15,802	2,488	2.34	0.00
NYC	583,604	18,609	8,266	3.50	0.00

Table 3: Summary statistics for different regions, along with power-law parameters with p-values. x stands for place ranks for all transitions within the given region, while x_{min} informs the x value where power-law scaling begins with α as the scaling exponent and x_{max} refers to the maximum place rank observed in the dataset. Statistically significant p-values are shown in **bold**.(i.e., no statistical significance is found for the power-law fit)

4.2 Goodness-of-Fit Tests

Before claiming whether power-law is a plausible fit for a given dataset, it is crucial to perform goodness-of-fit tests, in addition to statistically comparing the power-law with alternative heavy-tailed distributions [14]. In this section, we take a critical look at the power-law assumption for rank-distance distribution in the context of human mobility.

Instead of pursuing a purely qualitative analysis of the dataset (e.g., based on visual inspection), we quantitatively test the power-law hypothesis. We estimate the goodness-of-fit of power-law distribution using the KS statistic, to measure the similarity (or differences) between our dataset and hypothesized power-law model. Rather than describing the method here, we refer the readers to [3] (section 4, page 675) for a detailed statistical explanation. We wish to highlight the interpretation of p-value which is obtained as a result of the goodness-of fit test: p-value has an inverse interpretation, compared with significance testing – the higher the p-value, the higher the chance of observed data to follow a power-law like distribution, and vice-versa.

Tables 2 and 3 list p-values for a plausible power-law model for the three regions under study. In all the experiments in Section 4.1.1 and 4.1.3 (i.e., with or without an estimated lower bound), we have obtained a p-value of 0, implying that the likelihood of the power-law model to fit the observed place ranks is negligible.

The question thus arises is: if the place-rank distribution does not follow a power-law model, which other heavy-tailed distributions can better describe them? We investigate this issue in the next section.

4.3 Alternative Models

Now that we have obtained statistical evidence to suggest that place transitions do not follow a power-law model as a function of rank, we turn our attention towards finding competing distributions, if they exist, which can possibly explain a better fit. Our goal is to find a good model, that is a model which can explain our data well, as opposed to an "ideal" model.

To compare power-law nature of rank-distance distribution, we have chosen two similar heavy-tailed distributions which are listed below:

1. **Log Normal**, parameterized by μ and σ takes the form,
$$f(x) = \frac{1}{x}\exp\left[-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right]$$
2. **Power Law with Exponential Cutoff**, parameterized by the scaling exponent α_E and decay rate λ. Mathematically, it is described as:
$$f(x) = x^{-\alpha_E}e^{-\lambda x}$$

Note that a pure power-law is the limiting case of cutoff power-laws, which arises when $\lambda \to 0$.

In Section 4.1, Figures 1, 2, 3 and 4 illustrate the fit of these alternate models to the data. Furthermore, we have performed statistical tests to compare power-law with the above distributions. Table 4 lists the parameters for respective distributions for ZRH, CHE and NYC. In addition, it also reports log-likelihood ratios (LR) for both distributions in comparison with a pure power-law. Note that the comparison with the power-law distribution has been performed with x_{min} set to 1, i.e., for the complete dataset.

Note that LR with a negative sign favors the competing distribution over a pure power-law, while the p-value indicates its statistical significance. Higher the p-value, higher is the chance that observed sign of LR is a result of statistical fluctuations and thus the alternative distribution hypothesis can be rejected. (It is important to observe that the interpretation of p-value in this section differs from the interpretation described in Section 4.2.) From Table 4, it is clear that log-normal and cutoff power-laws are preferred over pure power-law model due to significant p-values ($p < 0.05$). We have obtained identical results for all three geographical regions, highlighting the fact that rank-distance distributions are better explained using cutoff power-law model as opposed to a pure power-law model.

5 Discussion and Conclusion

Modeling and analyzing human movement patterns have been an area of research and active debate in the scientific community. Various models of human mobility have been proposed in the literature. In recent times, due to the availability of large-scale spatial datasets obtained from location-based services like Foursquare, it has become possible to empirically validate some of these models at a scale, which was not feasible earlier.

In this paper, using data obtained from these services, we have taken a critical look into the power-law hypothesis of the rank-based model to characterize human mobility. We have found that the rank-distance distribution does not follow a pure power-law on an independently collected Foursquare data of a country (Switzerland), canton (Zurich) and a major metropolitan (New York City). Instead, we have observed that the rank-distance can be better explained using a power-law with exponential cutoff model, as opposed to a pure power-law model. We have performed the statistical analysis on this dataset and found results to be consistent.

We wish to highlight that even though we have observed the cutoff power-law parameter α_E, to be consistent across the three studied regions with values in the range of $0.86 - 0.93$ (Table 4), we do not claim a cutoff power-law model as the "universal" mobility model to explain human transitions. Furthermore, we clearly do not imply that these results hold true for other datasets from which human movement trajectories can be inferred such as cellular data records [5], GPS traces obtained from taxicab movements [11], etc. This has to be empirically verified and will be investigated as part of the future work.

Our study suggests that urban mobility patterns are more nuanced than previously reported in the literature and that rigorous statistical analysis including goodness-of-fit tests should to be performed in view of the generality of human mobility models.

6 Acknowledgments

We sincerely thank Trinh Minh Tri Do (Idiap) for his valuable insights and meaningful discussions. This work was funded by the SNSF HAI project.

	Log Normal				Power Law with Exp. Cutoff			
	LR	p	μ	σ	LR	p	α_E	$\lambda\ (\times 10^{-4})$
ZRH	-3,330.26	**0.00**	4.09	3.31	-6,294.633	**0.00**	0.86	2.4
CHE	-10,471.73	**0.00**	3.97	3.37	-16,570.95	**0.00**	0.93	1.2
NYC	-60,844.73	**0.00**	3.39	5.01	-198,489.1	**0.00**	0.91	0.40

Table 4: Distribution parameters for log normal and power law with exponential cutoff models. Log-likelihood ratios (LR) are also shown along with their respective p-values. Statistically significant p-values are shown in **bold**.

References

[1] Brockmann et al., D. The scaling laws of human travel. *Nature 439*, 7075 (2006), 462–465.

[2] Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. Exploring millions of footprints in location sharing services. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).

[3] Clauset et al., A. Power-law distributions in empirical data. *SIAM review 51*, 4 (2009), 661–703.

[4] Foursquare. https://foursquare.com/about, 2013. [Online; accessed May, 2013].

[5] Gonzalez et al., M. C. Understanding individual human mobility patterns. *Nature 453*, 7196 (2008), 779–782.

[6] Malmi et al., E. Checking in or checked in: comparing large-scale manual and automatic location disclosure patterns. In *MUM*, ACM (2012).

[7] Massey Jr et al., F. J. The kolmogorov-smirnov test for goodness of fit. *JASA 46*, 253 (1951), 68–78.

[8] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., and Mascolo, C. A tale of many cities: universal patterns in human urban mobility. *PloS one 7*, 5 (2012), e37027.

[9] Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011).

[10] Ravenstein, E. G. The laws of migration. *Journal of the Statistical Society of London 48*, 2 (1885), 167–235.

[11] Santani, D., Balan, R. K., and Woodard, C. J. Spatio-temporal efficiency in a taxi dispatch system. *6th International Conference on Mobile Systems, Applications, and Services, MobiSys* (2008).

[12] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science 327*, 5968 (2010), 1018–1021.

[13] Stouffer et al., S. A. Intervening opportunities: a theory relating mobility and distance. *American sociological review* (1940), 845–867.

[14] Stumpf et al., M. P. Critical truths about power laws. *Science 335*, 6069 (2012), 665–666.

[15] Wikipedia. http://en.wikipedia.org/, 2013. [Online; accessed May, 2013].