Heterogeneity of Device Contact Process in Pocket Switched Networks *

Ye Tian[1,2] and Jiang Li[3]

[1] Anhui Province Key Laboratory on High Performance Computing,
[2] School of Computer Science and Technology,
University of Science and Technology of China, Hefei, Anhui 230026, China
yetian@ustc.edu.cn
[3] Department of Systems and Computer Science,
Howard University, Washington DC 20059, USA
lij@scs.howard.edu

Abstract. Understanding device pair's contacts is essential in pocket switched networks (PSN). However, most of the studies on this issue are focused on the empirical distribution aggregating inter-contact times from all the device pairs, and seeking to find common characteristics of their contact processes. In this paper, we present an insightful analysis on both the aggregated and the pair-wise inter-contact times obtained from three real-world datasets. We find that device pairs are heterogeneous in many aspects, including not only their contact frequencies, but also their contact patterns. More surprisingly, even for those frequently contacting pairs, their behaviors are diverse, and could not be described with a universal model. Finally, implication of the observed heterogeneity on PSN's message forwarding algorithm is discussed, and we show that with the awareness of the device pair's heterogeneous contact pattern, the network's message relaying service could be improved considerably.

Key words: Delay tolerant networks, pocket switched networks, inter-contact times, message forwarding algorithm

1 Introduction

With the advance of wireless technologies and prevalent use of portable wireless devices, in recent years, the idea of pocket switched network (PSN), which is a special case of the delay tolerant network (DTN), has been proposed (e.g. the Haggle project [1]). In PSN, portable wireless devices such as cell phones and PDAs carried by human beings form an ad-hoc network. In such a network, contacting among devices is the only opportunity for communication, therefore,

* This work was funded by the Specialized Research Fund for the Doctoral Program of Higher Education of China 20093402120020, and was also funded in part by US NSF grant CNS-0832000 and the Mordecai Wyatt Johnson Program at Howard University.

it is highly important for people to understand the device pair's contact pattern, especially the intervals between their consecutive contacts (referred to as *inter-contact times*). Previous studies [2][3] on this topic are mainly based on the empirical distribution aggregating inter-contact times of all the device pairs, with an assumption that the aggregated distribution could represent individual pair's contact process. However, [4] suggests that device pairs are in fact heterogeneous regarding their contacting behavior. In this work, we study both the aggregated inter-contact time distribution in percentiles and the individual pair's distribution.Three real-world datasets with mobility and contact information of large numbers of portable wireless devices are exploited to identify and understand the heterogeneity of device pairs' contact processes. Finally, we discuss the implication of the observed heterogeneity on the design of the PSN's message forwarding algorithms.

The remainder part of this paper is organized as follows: in Section 2, related works are surveyed and their relationships with our work are discussed; in Section 3, real-world datasets are analyzed, in particular, the aggregated distributions of the device pairs' inter-contact times are studied in percentiles; we investigate and classify frequently contacting pairs, and give some interpretations for the heterogeneity observed in Section 4; in Section 5, the implication of our observation is discussed; finally in Section 6, we conclude this paper and discuss the future work.

2 Related Work

Our work focuses on studying the contact pattern of device pairs in PSN networks, and we are especially interested in pair's inter-contact time between consecutive contacts. Previous studies on this topic are mostly based on the empirical distribution aggregating the inter-contact times from all the device pairs. In [2], the authors study aggregated distributions from a number of real-world datasets, and find that the inter-contact time follows power law and could be modeled with a truncated Pareto distribution. This finding contradicts the assumption in many works (e.g. [5]) that a pair's contact process is Poisson with exponential inter-contact times. In [3], it is reported that a dichotomy exists in the aggregated inter-contact time distribution: the distribution is power law in certain range, but it has an exponential tail. A random walk model with infinite sites is used to explain the observed power-law inter-contact time. While in [6], a random walk model in an unbound domain is applied for the same purpose. This work differs with these works in that we only study the inter-contact time distribution of an individual pair or the aggregated distribution of pairs in a small group, and we interpret our findings with well-founded theories and widely recognized observations.

Only a few works address pair-wise inter-contact times. In [7], the authors use the Dartmouth dataset for analyzing device pair's contact process, and draw a conclusion that for majority of the device pairs, their inter-contact times are exponentially distributed. In [4], it is shown that the mean inter-contact time

is heterogeneous, and the inter-contact times for most of the pairs could be described with a lognormal distribution. Our work differs from these previous works in that we classify the pairs based on their contact patterns and only consider the distributions backed up with theoretical or experimental supports for describing their inter-contact times, and we also discuss the implication of our observations on the design of PSN's message forwarding algorithm.

3 A First Look at the Contact Process

For our study, we select three datasets containing long-time mobility and contact information of a large number of wireless devices, which are the dataset from the MIT Reality Mining project [8], the Dartmouth dataset [9], and the dataset from UCSD Wireless Topology Discover project [10]. For the remainder part of this paper, we simply refer to them as Reality, Dartmouth, and UCSD respectively. Among the three datasets, devices contacts were recorded in Reality, while for Dartmouth and UCSD, we considered a contact between two devices happened if they were associated with a same AP simultaneously, as assumed in the previous works[2][3].

Table 1. Pairs and contacts in different percentiles

Percentile	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Reality	26.77%	23.88%	15.98%	11.01%	7.36%	5.18%	3.57%	2.93%	2.06%	1.26%
Dartmouth	0.79%	15.30%	32.86%	21.70%	12.78%	6.71%	4.72%	2.67%	1.61%	0.84%
UCSD	3.83%	23.90%	22.40%	15.25%	9.31%	7.64%	4.55%	4.62%	4.87%	3.63%

By investigating the three datasets, we are trying to answer the question that whether or not the contact processes for all the device pairs could be viewed homogenous, and if not, in what aspects the heterogeneity exists. For this purpose we study the aggregated distribution of the pairs' inter-contact times. However, unlike previous works (e.g., [2] and [3]), we do not take all the device pairs into consideration, but group them according to their contact frequencies and study inter-contact times within each group. Concretely, we sort all the pairs in an ascending order regarding their mean inter-contact times, and group the pairs in each ten percentiles. For example, by denoting the percentile group of "10-20", we mean the pairs with their mean inter-contact times between the first 10 and the first 20 percents among all the pairs in this order.

We list in Table 1 the percentage of the contacts made by each group for the three datasets. One can see that in these datasets, some percentile groups make much more contacts than other groups. Moreover, it is observed that the contact number is decreasing very sharply in groups with lower percentiles for each dataset, but the decreasing becomes smooth in higher percentiles. This observation indicates that we could roughly categorize these device pairs into frequently contacting pairs and infrequently contacting ones, and majority of

the contacts are made by the former. For example, pairs in the three most contacting percentile groups contribute 66.63%, 69.86%, and 61.55% of the total contacts in Reality, Dartmouth, and UCSD respectively.

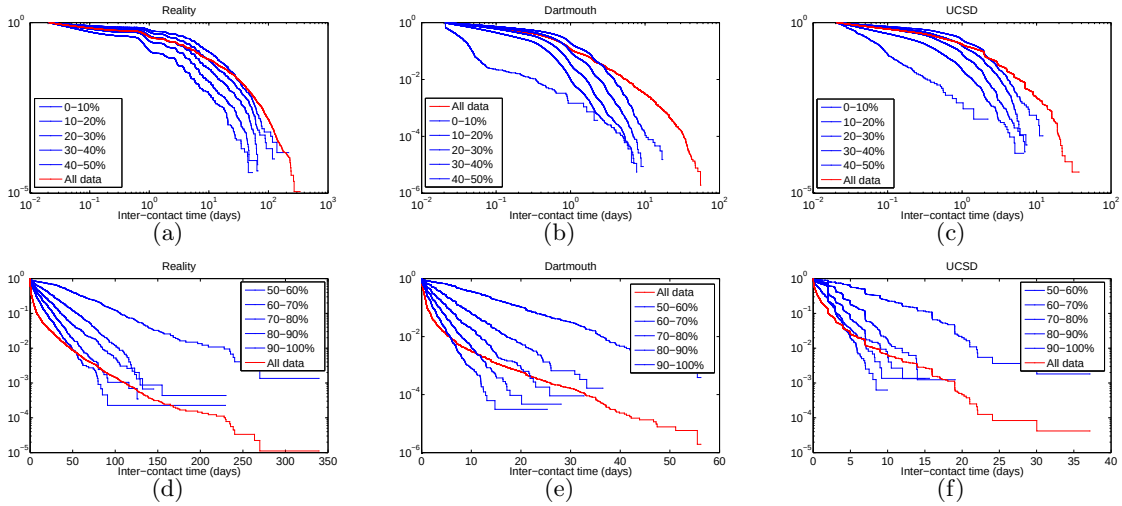

Fig. 1. Aggregated distributions of inter-contact times for pairs in 0-10, 10-20, 20-30, 30-40, and 40-50 percentiles in (a) Reality, (b) Dartmouth, and (c) UCSD; aggregated distributions of inter-contact times for pairs in 50-60, 60-70, 70-80, 80-90, and 90-100 percentiles in (d) Reality, (e) Dartmouth, and (f) UCSD.

In Figure 1, we plot the aggregated inter-contact time distribution in complementary cumulative distribution function (CCDF) for pairs in each percentile group of the three datasets. Please note that for percentiles between 0-50, we use the log-log scale, as in Figure 1(a-c), while for percentiles between 50-100, we use the linear-log scale, as in Figure 1(d-f). We also plot the aggregated distribution of all the pairs for comparison. From these figures, we find that for pairs in different percentiles, they differ not only in their contact frequencies, but also in the shape of the empirical distributions of their inter-contact times: From Figure 1(d-f) one could see that the distribution curves are approximated straight lines in the linear-log scale, suggesting that the inter-contact times are exponential. On the other hand, Figure 1(a-c) show that the aggregated distributions of the inter-contact times have near straight-line curves within the delays shorter than one day under the log-log scale, indicating that the distributions are more power-law like. The observation of the distribution shapes suggests that pairs in different percentile groups may have heterogeneous contact patterns in addition to their contact frequencies.

Summarizing our findings from Table 1 and Figure 1, we can conclude that the device pairs in all the three datasets are highly heterogeneous in at least two aspects: first, we find that contact frequencies of the device pairs are heteroge-

neous, for example, Table 1 shows that more than 60% of the contacts are made by only 30% of the most contacting device pairs; more importantly, we find that pairs have different contact patterns, where the inter-contact times for pairs in lower percentiles are more power-law like, while the inter-contact times for pairs in higher percentiles are more exponential. The observed heterogeneity suggests that it may be inappropriate to use a universal contact process model based on the aggregated inter-contact time distribution for interpreting all the pairs' behaviors. On the other hand, as we have seen that a small part of frequently contacting pairs contribute majority of the contacts, it is highly important for people to understand these pairs. With focus narrowed on these frequently contacting pairs, a natural question arises as whether or not these pairs' behaviors are same. We will answer this question in the next section.

4 Contact Process of Frequently Contacting Pair

4.1 Statistical analysis

In this section, we focus on the frequently contacting device pairs and try to understand their behaviors. For the first step of our study, we filter out those infrequently contacting pairs with a threshold. That is, we only consider the pairs with their contacts more than the threshold. We then use the Cramer-Smirnov-Von-Mises test[11] to study the frequently contacting pairs filtered out. The Cramer-Smirnov-Von-Mises test is a statistical method to testify whether or not the sampled data is compatible with a given distribution function. In our test, the sampled data is the inter-contact times of the pair under study, and we consider the following candidate distribution functions.

- Exponential distribution: For this candidate, we use a CDF function as

$$F(x) = 1 - e^{-\lambda_{ij} x}, x \geq 0$$

  where λ_{ij} is a constant contact rate between the two devices of the pair.
- Pareto distribution: For this candidate, we use a CDF function as

$$F(x) = 1 - \left(\frac{x}{\beta_{ij}}\right)^{-\alpha_{ij}}, x \geq \beta_{ij}$$

  where β_{ij} is the minimum inter-contact time observed for this pair and α_{ij} is the slope rate of the empirical CDF curve in the log-log scale.

For obtaining the parameter λ_{ij} of the empirical exponential distribution, we simply let λ_{ij} be the inversion of the mean inter-contact time. While for the parameters of the empirical Pareto distribution, we use the maximum likelihood estimator of α_{ij} [12] as

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln \frac{x_i}{\beta}}$$

where n is the number of the samples and x_i is the ith data sample.

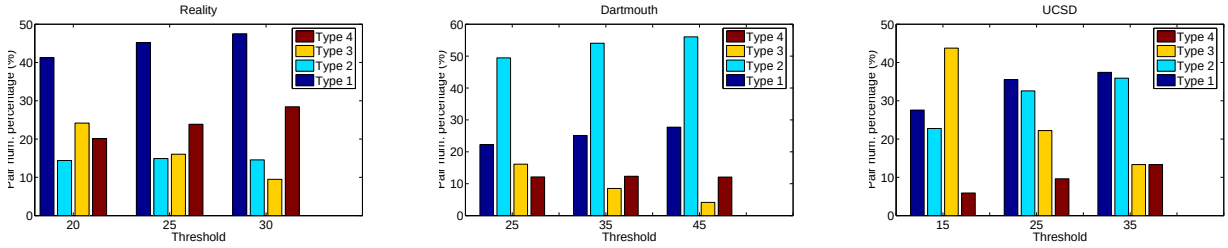

Fig. 2. Histograms of slope rates for Type 1 pairs in (a) Reality, (b) Dartmouth, and (c) UCSD

The reason that we choose the exponential and Pareto distributions is that they are either supported by well-known theory or widely observed in similar activities. For the exponential distribution, it is well known that if a node's location process is independent, stationary and ergodic, the long-term contact process between any two nodes is Poisson[5]. Moreover, it is proved that the exponential inter-contact time could be generated with popular mobility models such as the Random Way Point model and the Random Direction model [13]. On the other hand, recent studies show that for many human activities, the inter-event time is heavy-tailed and follows a Pareto distribution. These activities include composing emails[14], visiting websites[15], responding surface mails[16], and performing financial transactions[17]. Although human contacting is different from these activities, however, it is highly possible that there are some similarities, therefore we also consider the Pareto distribution as our candidate distribution function.

The Cramer-Smirnov-Von-Mises test is a statistical method which compares the sampled data with the hypothetic candidate distribution, and based on a rejection level α, it returns a result of positive or negative on whether or not the sampled data is compatible with the candidate distribution. Here the rejection level α is the probability of the test to make false positive errors. As we have two candidate distributions, there are two tests for each pair, which are referred to as the exponential test and the Pareto test respectively, and for each test we will have two results. We categorize the pairs into four types based on their testing results:

- Type 1: pairs passing only the Pareto test;
- Type 2: pairs passing only the exponential test;
- Type 3: pairs passing both tests;
- Type 4: pairs passing none of the tests.

As we are trying to categorize pairs' behavior patterns, the rejection level α must be selected carefully: if α is set too large, the tests are very loose, and many pairs will pass both tests, suggesting that we are actually failed to differentiate them; if α is too small, the tests are very selective, and we may find many pairs failing to pass any test just because of the raw nature of the data samples. In our

study we choose $\alpha = 0.01$ for Reality and Dartmouth, and $\alpha = 0.05$ for UCSD. We also change the threshold of the contact number for each dataset for filtering out infrequently contacting pairs, and show the test results in percentages of the pairs falling in each type for the three datasets in Figure2. From the figure, one can see that for each dataset, when the threshold is small, there are considerable numbers of pairs falling in all the four types; however, when the threshold gets increased, Type 3 pairs are filtered out rapidly. Recall that Type 3 pairs are the pairs passing both tests, it is reasonable to consider these pairs as infrequently contacting pairs, as they pass both tests simply because there is no sufficient sampled data for the tests.

We then focus on the other three types, i.e., Type 1, 2, and 4. From the figure one can see that each type has a persistent portion under test, and the pairs in the three types are not easily filtered out by increasing the threshold. Based on this observation, we could consider the Type 1, 2, and 4 pairs as representative frequently contacting pairs. In other words, we could categorize the frequently contacting device pairs into the three types, i.e. pairs with exponential inter-contact time, pairs with Pareto inter-contact time, and pairs not belonging to the above two.

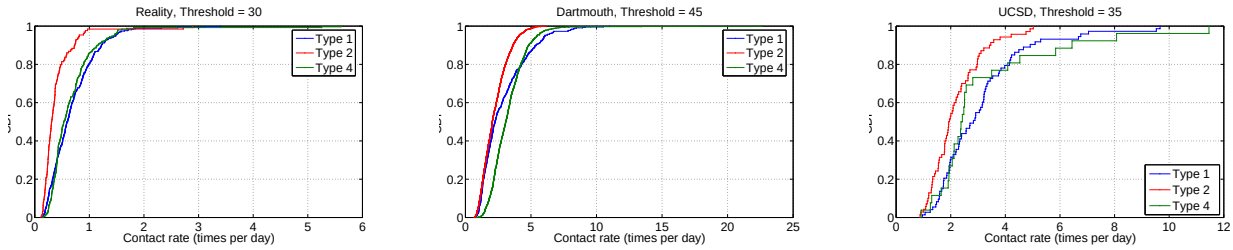

Fig. 3. CDFs of the contact rates for Type 1, 2, and 4 pairs in (a) Reality, (b) Dartmouth, and (c) UCSD

To have a further understanding of the Type 1, 2, and 4 pairs, we also estimate a pair's contact rate by averaging its inter-contact times and taking the inversion. The distribution of the contact rates for each type of all the three datasets are plotted in Figure 3. From the figure one can see that although pairs of all the three types are considered as frequently contacting, there are still some differences: Type 1 and Type 4 pairs contact obviously more frequently than Type 2 pairs. We also study the empirical Pareto distribution's slope rate for Type 1 pairs in the three datasets, and find that all the slope rates are smaller than 1 and are concentrating around $0.3 \sim 0.5$.

4.2 Discussion and interpretation

From the above statistic test results, it is implied that there are at least two contact patterns for the frequently contacting pairs, i.e. the contact pattern producing Pareto inter-contact time and the one producing exponential time. As shown in [13], exponential inter-contact time could be caused by device's independent, stationary, and ergodic location process, therefore we mark these contacts as "unintended", as this kind of contact is a byproduct of a node's independent visiting to some locations. For example, neighbors at home or working place may often make "unintended" contacts, but people do not make these contacts on purpose. For this reason, we refer to the relationship between the two parties of a pair producing exponential inter-contact time as "familiar strangers".

While for the Pareto inter-contact time, we believe they are caused by similar reasons as the Pareto inter-event time observed in many human activities, such as the task priority[14] and the human interest[18]. In other words, we believe that the Pareto inter-contact times could also be explained with objective reasons of human beings, as human has the interest or urgency to perform the contact on purpose. Therefore we could mark the contacts with Pareto inter-contact time as "intended" contacts and could refer to the relationship between the two parties of such pair as "friends". Finally, for Type 4 pairs, we explain that these pairs are both "familiar strangers" and "friends", and their contacts are a combination of the two types of the contacts, but none of them dominates. Finally, we observe that Type 1 pairs contact more frequently than Type 2 pairs, which conforms to the intuition that close friends meet more often than familiar strangers.

5 Implication on PSN Message Forwarding

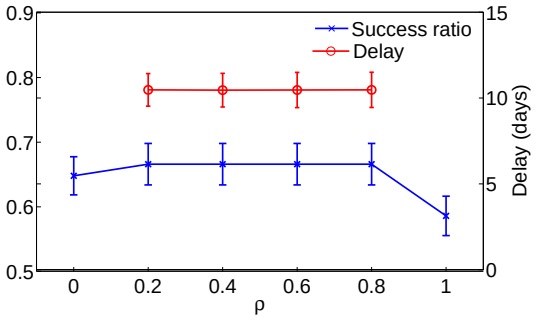

Fig. 4. Success ratio and delay for PSN message forwarding

From the statistical studies in the above section, we find that frequently contacting device pairs may have their inter-contact times following exponential or Pareto distribution. It is well known that for exponential inter-contact times,

a pair's contacting frequency indicates its expected next contacting time, due to the memoryless property of the exponential distribution. However, for the Pareto inter-contact times, as the distribution has memory, a device pair that has just contacted is likely to make a contact in near future, while a pair that contacted long ago is not likely to make a contact very soon. Therefore, to predict how soon a pair of devices may contact, we should consider the last recent contact age for the Pareto inter-contacting device pairs as well as the contact frequency for the exponential inter-contacting one.

To testify our point, we simulate a PSN network using the UCSD dataset, and examine the success ratios and delivery delays of the network's message delivery jobs when nodes apply different forwarding strategies. Figure 4 shows that when both the node pair's contact frequency and the last contact age are considered (corresponding to $0 < \rho < 0$ in the figure), performance of PSN's message delivery job is better than the cases when only the contact frequency (corresponding to $\rho = 0$ in the figure) or the last contact age (corresponding to $\rho = 1$ in the figure) is concerned. Please note that when $\rho = 0$, it is exactly the "FRESH" algorithm proposed in [19], and for $\rho = 0$, it is the "Greedy" algorithm studied in [20].

Moreover, as we have categorize the node pairs into "familiar strangers" and "friends", and classify the contact processes into "intended" and "unintended", if the relationship of the node pair, or even more, the type of the contact process is available, PSN's message forwarding algorithms should make much more accurate predictions in selecting the next hop for message forwarding. Unfortunately, although we have statistically identified the types for some node pairs and contact processes of the datasets studied in this work, precisely knowing the exact type of each node pair and contact process is still infeasible using any existing datasets, therefore we are unable to examine the exact benefits brought by the awareness of the node pair relationship and contact process type experimentally.

6 Conclusion and Future Work

In this paper, we study three datasets containing contact and mobility information of wireless devices, for understanding the device contact process under the context of PSN networks. We first group pairs into different percentiles based on their contact frequencies and study their aggregated inter-contact time distributions. We find that pairs are heterogeneous in many aspects, including their contact frequencies as well as their contact patterns. We then study the pair-wise inter-contact times, and find that even for the frequently contacting pairs, they are behaving diversely. We categorize the frequently contacting pairs into three types by using the Cramer-Smirnov-Von-Mises test, and apply different theories for interpreting their different contact patterns. Furthermore, we discuss the implication of our observation on the contact processes for the PSN message forwarding algorithm, and show that with the awareness of the heterogeneity, better performance on PSN's message delivery should be expected.

Our future work is in two directions. First, more insightful studies are required to meet the gaps between the empirical observation and the theoretical model, such as explaining the Type 4 pairs. More importantly, we need to work on further exploiting the inferred contact pattern for solving critical problems, such as routing, security, resource management, and quality of service, on PSN networks. For example, "friends" type node pairs should be considered more reliable in some mission critical applications.

References

1. Haggle Project: http://www.haggleproject.org.
2. Chaintreau, A., Hui, P., Diot, C., Gass, R., Scott, J.: Impact of human mobility on the design of opportunistic forwarding algorithms. In: Proc. of IEEE INFOCOM'06, Barcelona, Spain (April 2006)
3. Karagiannis, T., Boudec, J.Y.L., Vojnovic, M.: Power law and exponential decay of inter contact times between mobile devices. In: Proc. of ACM MobiCom'07, Montreal, Canada (September 2007)
4. Conan, V., Leguay, J., Friedman, T.: Characterizing pairwise inter-contact patterns in delay tolerant networks. In: Proc. of ACM Conference on Autonomic Computing and Communication Systems, Rome, Italy (October 2007)
5. Grossglauser, M., Tse, D.N.C.: Mobility increases the capacity of ad hoc wireless networks. IEEE/ACM Trans. Networking **10**(4) (2002) 477–486
6. Cai, H., Eun, D.Y.: Crossing over the bounded domain: from exponential to power-law inter-meeting time in manet. In: Proc. of ACM MobiCom'07, Montreal, Canada (September 2007)
7. Conan, V., Leguay, J., Friedman, T.: The heterogeneity of inter-contact time distributions: its importance for routing in delay tolerant networks. arXiv:cs/0609068v2 [cs.NI] (January 2007)
8. MIT Reality Mining Project: http://reality.media.mit.edu.
9. CRAWDAD: http://crawdad.cs.dartmouth.edu.
10. UCSD WTD Project: http://sysnet.ucsd.edu/wtd.
11. Eadie, W.T.: Statistical Methods in Experimental Physics. Elsevier Science (1983)
12. Asrabadi, B.R.: Estimation in the pareto distribution. Physica **37**(1) (1990) 199–205
13. Groenevelt, R., Nain, P., Koole, G.: The message delay in mobile ad hoc networks. Perform. Eval. **62**(1-4) (2005) 210–228
14. Barabasi, A.L.: The origin of bursts and heavy tails in human dynamics. Nature **435** (2005) 207–211
15. Dezso, Z., Almaas, E., Lukacs, A., Racz, B., Szakadat, I., Barabasi, A.L.: Dynamics of information access on the web. Phys. Rev. E **73** (2006) 066132
16. Oliveira, J.G., Barabasi, A.L.: Human dynamics: the correspondence patterns of Darwin and Einstein. Nature **437** (2005) 1251
17. Plerou, V., Gopikrishnan, P., Amaral, L.A.N., Gabaix, X., Stanley, H.E.: Economic fluctuations and anomalous diffusion. Phys. Rev. E **62** (2000) 3023–3026
18. Han, X.P., Zhou, T., Wang, B.H.: Modeling human dynamics with adaptive interest. New J. Phys. **10** (2008) 073010
19. Dubois-Ferriere, H., Grossglauser, M., Vetterli, M.: Age matters: efficient route discovery in mobile ad hoc networks using encounter ages. In: Proc. of ACM MobiHoc'03, Annapolis, MD, USA (June 2003)

20. Erramilli, V., Chaintreau, A., Crovella, M., Diot, C.: Diversity of forwarding paths in pocket switched networks. In: Proc. of ACM IMC'07, San Diego, CA, USA (October 2007)