1

A complexity science perspective on human mobility

Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, Dashun Wang

Fueled by big data collected by a wide range of high-throughput tools and technologies, a new wave of data-driven, interdisciplinary science have rapidly proliferated during the past decade, impacting a wide array of disciplines, from physics and computer science to cell biology and economics. In particular, the ICT's are inundating us with huge amounts of information about human activities, offering access to observing and measuring human behavior at an unprecedented level of details. These large-scale datasets, offering objective description on human activity patterns, have started to reshape, and are expected to fundamentally alter, our discussions on quantifying and understanding human behavior. An impressive shift has been witnessed in statistical physics and complex system theory since the beginning of the new millennium, when the possibility of analyzing large datasets of human activities and social interactions has boosted a renewed interest in the study of human mobility on one side, and of social networks on the other side.

The understanding of how objects move, and humans in particular, is a longstanding challenge in the natural sciences, since the seminal observations by Robert Brown in the 19th century, but it has attracted particular interest in recent years, due to the data availability and to the relevance of the topic in various domains, from urban planning and virus spreading to emergency response. A first contribution of this chapter is to provide a brief account of this body of research, with a focus on the recent results on the empirical laws that govern the individual mobility patterns: we discuss how the key variables of people's travels (such as length, duration, radius of gyration, ...) follow universal laws, validated against different datasets of real observations. We also discuss how predictable people's movements are, illustrating recent findings indicating that the high degree of predictability of human motion is a universal

characteristic of every individual, despite the wide variety of individual whereabouts.

Next, we move from individuals to interactions — links — among individuals, and enter the domain of social network analysis. An extraordinary effort has been devoted to understanding the interconnectedness of individuals, i.e., the structure of the social networks we inhabit, and how this structure influences social phenomena, such as the importance of certain individuals or groups, the diffusion of information or the formation of communities. The second contribution of this chapter is to provide a brief account of the key findings of network science so far (what are the distinctive features of real social networks compared to random networks, how the community structure of real networks models the fabric of society, what are the mechanistic processes that generates realistic networks), to the purpose of discussing the recent results on how human mobility shapes and impacts social relations, and the other way around. Again, empirical laws were found that offer quantitative accounts of the intuition that people from the same social circles tend to co-locate in space and time more than people that are far apart in the social network. Building on this relation among social and mobility variables, it is possible to shed more light on how social networks (and mobile behavior) evolves over time.

We believe that the results surveyed in this chapter, about individual mobility laws and the relations between social ties and mobility, should become basic tools for research in various disciplines, and we envisage that the convergence between data mining research and network science research, already apparent in some of the works discussed here, will represent a strong trend in the near future, aimed at combining the analytical power of statistical physics and knowledge discovery.

1.1 Models of human mobility

We live in an era in which understanding individual mobility patterns is of fundamental importance for epidemic preventions and urban and transportation planning. Yet, human movements are inherently massive, dynamical, and complex. Indeed, on one hand, aided by modern transportation technologies, we can now travel to any place on the globe in just a day or two. On the other hand, while the mobility of our fellow species is mainly governed by mating needs and food resources, human mobility is fundamentally driven by ourselves, from job-imposed restric-

tions and family related programs to involvement in routine and social activities. Therefore, quantifying the regularities and singularities behind human movements had remained as a often elusive goal. Thanks to the availability of large-scale datasets generated by various domains of modern technologies, ranging from registration of dollar bills to mobile phone services and GPS devices to location based websites, we have witnessed a proliferation of studies on human mobility.

In this section, we will start from the most fundamental models for motions, rooting back to the 19th century. We will then describe several empirical observations of human mobility and the new generation of mobility models, presenting to what extent real human mobility patterns deviate from those expected from simple diffusion processes.

1.1.1 Motion models: Brownian motion and Lévy flights

In 1827, while he was studying sexual relations of plants, botanist Robert Brown noticed that granules contained in grains of pollen were in constant motion, and that this motion was not caused by currents in the fluid or evaporation. He thought at first that they were jiggling around because they were alive or because of the organic nature of the matter. So, he did the same experiment with dead organic and inorganic matter finding there was just as much jiggling. The movement evidently had nothing to do with the substance ever being alive or dead, and this left him and his contemporaries with a puzzling question: What is this mysterious perpetuum motion that keeps the pollen moving?
A possible explanation for the so-called **Brownian motion**[a], is that all the molecules in the fluid are in vigorous motion, and these tiny granules are moved around by this constant battering from all sides as the fluid molecules bounced off. Imagine we are in the middle of a crowd and there is a big balloon. As the individuals move around, they push the balloon from all directions: sometimes the balloon will move to the left, occasionally to the right, overall displaying a random, jittery motion like paths in Figure 1.1. A particle of pollen behaves like a really huge balloon in the midst of a dense crowd.

Such atomic-molecular thesis was guessed by Einstein, who in 1905 published a theoretical analysis of Brownian motion and showed that the mean distance reached by particles from the first collision point must

[a] The first observation of Brownian motion was reported in 1785 by the Dutch physician Jan Ingenhaysz. However, Brown was the first to discover the ubiquity of the phenomenon.

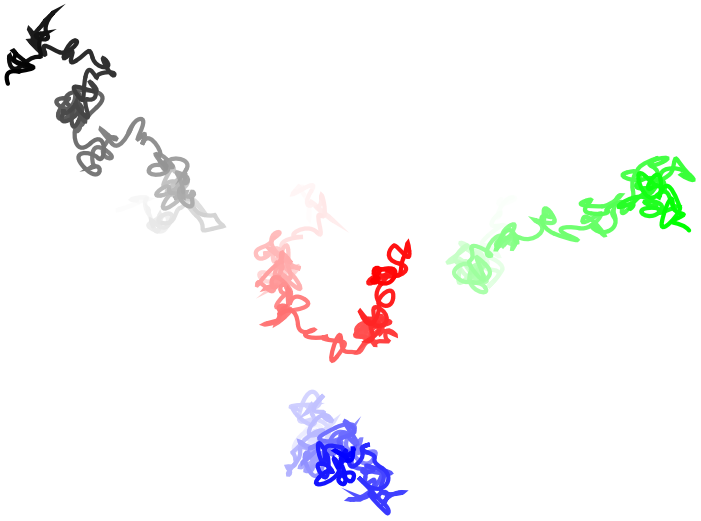

Figure 1.1 Some examples of Brownian motions.

grow with the square root of time. It means, for example, that after 4 seconds, the distance is only twice ($\sqrt{4} = 2$) the one found after a second, and not four times as insight would suggest. Einstein's calculations were confirmed experimentally in 1908 by physicist Jean Baptiste Perrin, who convinced even the most skeptical about the validity of atomic-molecular hypothesis.

Before Einstein, Louis Bachelier derived independently several mathematical properties of Brownian motion, including the equation for the probability $P(x,t)$ for the position x of a Brownian random walker at time t, when the walker starts as the origin at time $t = 0$. The equation for $P(x,t)$ in one dimension is given by the *diffusion equation*, with a Gaussian solution. Therefore, a Brownian motion is basically a random walk with a normal distribution for the position of the random walker after a time t, with the variance proportional to t. It means that random walkers tend to travel roughly the same distance between sightings.

However, there are situations in which equations for Brownian motion are no longer applicable. An example occurs if the jumps are of very large distances: this is the case of some animal movements. Measurements on albatrosses, monkeys and marine predators, suggested that animal trajectories are different from the Brownian motion, and they are better approximated by the so-called **Lévy-flight**. The French mathematician Paul Lévy investigated in the 1930's the mathematics of random walks

with infinite moments. A random walk of N steps is a sum of N indipendent and identically distributed random variables with mean $\mu = 0$ and variance σ^2, that is $S_N = X_1 + X_2 + \ldots + X_N$. Lévy posed the following question: when the probability distribution $P_N(x)$ of the sum of N steps have a similar form as the probability distribution of a single step $p(x)$? For walks with finite jump variances, the central limit theorem implies that the overall probability $P_N(x)$ is a Gaussian. For infinite variance random walks, the Fourier transform of $p(x)$ has the form $\bar{p}(k) = e^{-|k|^\beta}$ with $\beta < 2$. The Gaussian distribution (Brownian motion case) corresponds to $\beta = 2$, and the Cauchy distribution corresponds to $\beta = 1$. Therefore Lévy-flights are a generalization of Brownian motions (Figure 1.2).

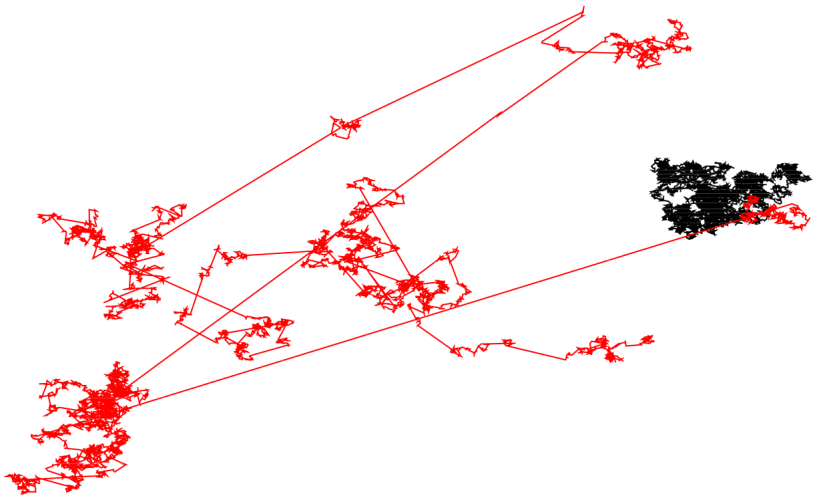

Figure 1.2 Brownian motion (black curve) is describe as a random walk in which all the steps give the same contribution. Lévy-flight (red curve) occurs when the trip is dominated by a few very large steps.

When the absolute value of x is large, $p(x)$ is approximately $|x|^{-(1-\beta)}$, which implies that the second moment of $p(x)$ is infinite when $\beta < 2$. This means that there is no characteristic size for the random walk jumps, except in the Gaussian case of $\beta = 2$. It is just this absence of a characteristic size the makes Lévy random walks scale-invariant fractals.

1.1.2 Human mobility patterns

Are human movements similar to those of grains of pollen, following a Brownian motion, or are they governed by Lévy-flight, like marine predators and monkeys? Or do they follow their own laws? To answer above questions, we need to observe humans under a microscope, like Perrin observed atoms and was able to experimentally confirm Einstein's theory. The technological era, at last, allows us to track human mobility and to test models, thanks to the exploding prevalence of mobile phones, GPS, and other handheld devices. Such devices are our social microscopes.

In 2006, Dirk Brockmann and his colleagues proposed using the geographic circulation of bank notes in the United States as proxy for human traffic, based on the idea that individuals transport money as they travel. They analyzed data collected at the largest online bill-tracking Website `www.wheresgeorge.com`, and found that most bills remain in the vicinity of their initial entry, yet a small but a significant number have traversed distances of the order of the size of USA (Figure 1.3), consistent with the intuitive notion that short trips occur more frequently that long ones. Brockmann's team calculated that the probability $P(r)$

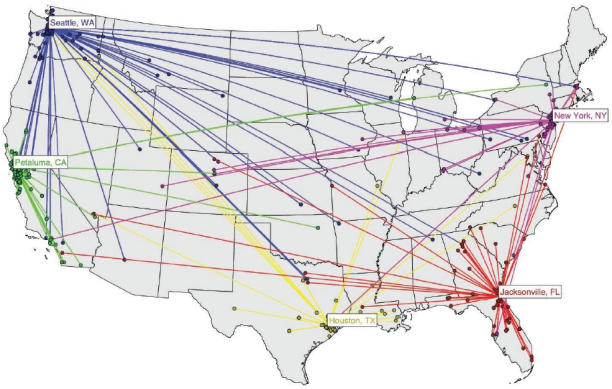

Figure 1.3 Short time trajectories of dollar bills in the United States. Lines connect origin and destination locations of bank notes that traveled for less than a week. Figure from Brockmann et al (2006).

of a bank note traversing a distance r follows a power law:

$$P(r) \sim r^{-(1+\beta)} \tag{1.1}$$

with an exponent $\beta \approx 0.6$. Moreover, they found that the typical distance

$X(t)$ from the initial starting point as a function of time is a power law:

$$X(t) \propto t^{1/\beta}. \tag{1.2}$$

As we know, for Brownian motion the distance $X(t)$ scales according to the square-root law. For a power law the variance diverge for exponents $\beta < 2$ and it implies that bank note dispersal lacks a typical length scale resembling Lévy-flights. Lévy-flights are superdiffusive; they disperse faster than ordinary random walks. This discovery was a major breakthrough in understanding human mobility on global scales. In light of this discovery, in dispersal humans are similar to animals.
However, our intuition suggests that we do not move completely random. There are regularities in our lives: most of us have a home, a work, an hobby. These activities necessarily shape our trajectories. Instead, if we do follow a pure Lévy flight we rarely find our way back home, but our position increasingly moves away from the initial one.

To further investigate human mobility patterns, in 2008 Barabási and his team analyzed the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period. Contrary to bills, mobile phones are carried by the same individual during his daily routine, offering the best proxy to capture individual human trajectories. An immediate result of the research was that the distribution of displacements Δr between user's positions at consecutive calls is well approximated by a truncated power-law:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} exp(-\Delta r/\kappa) \tag{1.3}$$

with exponent $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5$ km and some cutoff values κ. Such equation suggests that human motion follows a truncated Lévy-flight, apparently confirming in a certain way observations on bank notes. However, differences from randomness emerge from other measures. The distribution $P(r_g)$ of radius of gyration r_g, the characteristic distance traveled by a user when observed up to time t, also follows a power law, in contrast with random walks (Figure 1.4, left). So, most people usually travel in close vicinity to their home location, while a few frequently make long journeys. Furthermore, the probability $F_{pt}(t)$ that a user returns to the position where he was first observed after t hours shows several peaks at 24 h, 48 h and 72 h (Figure 1.4, right), capturing the recurrence and temporal periodicity inherent to human mobility.
The most important result was the finding that, after appropriate rescaling aiming to remove the anisotropy and the r_g dependence, all individuals seem to follow the same universal probability distribution $\tilde{\Phi}(\tilde{x}, \tilde{y})$

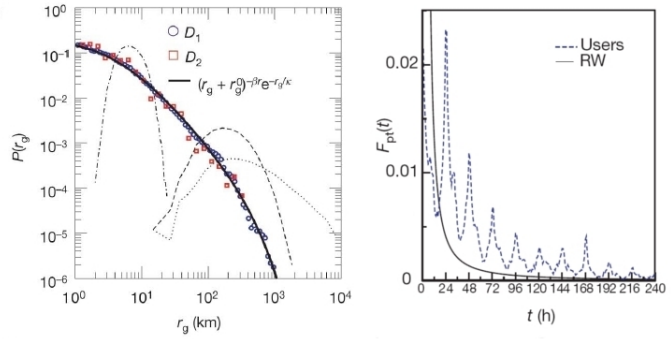

Figure 1.4 The distribution $P(r_g)$ of the radius of gyration measured for the users. The solid line represents a similar truncated power-law fit. The dotted, dashed and dot-dashed curves show $P(r_g)$ obtained from random walk, pure and truncated Lévy flights models. The picture on the right shows that the prominent peaks capture the tendency of humans to return regularly to the locations they visited before, in contrast with the smooth asymptotic behavior (solid line), predicted for random walks. Figure from González et al (2008).

that an individual is in a given position (x, y) (Figure 1.5 **b**). Individuals display significant regularity, returning to a few highly frequented locations, such as home or work. This regularity does not apply to the bank notes: a bill always follows the trajectory of its current owner; that is, dollar bills diffuse, but humans do not.

Song et al. extended the experiment to a larger dataset and measured the distribution of the visiting time (the interval Δt a user spends at one location). The resulting curve is well approximated by a truncated power-law with an exponent $\beta = 0.8 \pm 0.1$ and a cutoff of $\Delta t = 17$ h, witch the authors connected with the typical awake period of humans. The number of distinct location $S(t)$ visited by humans is sublinear in time, well approximated by $S(t) \sim t^{\mu}$ with $\mu = 0.6 \pm 0.02$, that indicates a decreasing tendency of people to visit previously unvisited locations. Moreover, the visitation frequency, that is the probability f of a user to visit a given location, is rather uneven, resulting in a Zipf-like visitation frequency distribution $P(f) \sim f^{-(1+1/\zeta)}$.

1.1.3 Predictability of human mobility

What is the role of randomness in human behavior and to what degree is human behavior predictable? This question is crucial, because the quan-

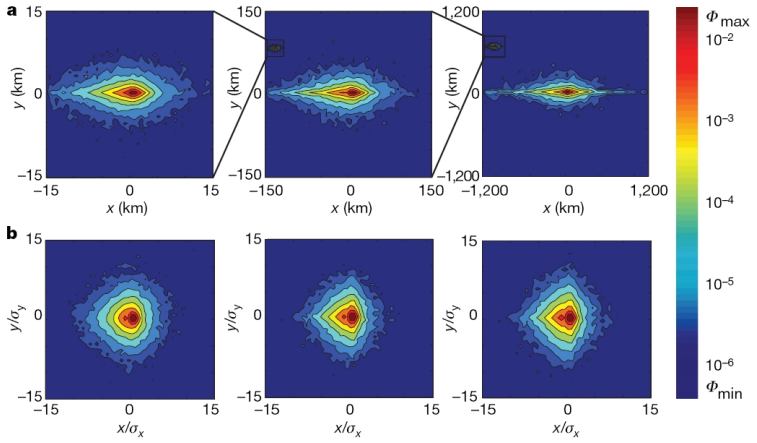

Figure 1.5 **a**, The probability density function $\Phi(x, y)$ of finding a mobile phone user in a location (x, y) in the user's intrinsic reference frame. The three plots, from left to right, were generated for $10,000$ users with: $r_g \leq 3$, $20 \leq r_g \leq 30$ and $rg > 100$ km. The trajectories become more anisotropic as r_g increases. **b**, After scaling each position, the resulting probability distribution has approximately the same shape for each group. Figure from Song et al (2009).

tification of the interplay between the predictable and the unforeseeable is very important in a range of applications. From predicting the spread of human and electronic viruses to city planning and resource management in mobile communications, our ability to foresee the whereabouts and mobility of individuals can help us to improve or save human lives.

In 2009, Song *et al.* provided a quantitative evaluation of the limits in predictability for human walks, using a 3-month-long mobile phone dataset of about $50,000$ individuals. The authors defined three entropy measures: the random entropy S_i^{rand} in the case of location visited with equal probability; the entropy S_i^{unc} that depends only on frequencies of visits; and the real entropy S_i that considers the probability of finding particular time-ordered subsequences in the trajectory. To characterize the predictability across the user population, they determined these three entropies per each user i, and calculated the distributions $P(S_i^{rand})$, $P(S_i^{unc})$ and $P(S_i)$, i.e. the frequency of entropy values. As shown in Figure 1.6**A**, $P(S_i)$ has a peak in $S = 0.8$ indicating that the real uncertainty in a typical user's whereabouts is $2^{0.8} \approx 1.74$. It means that a user who chooses randomly his or her next location could be found

on average in two locations. A big difference emerges in respect to the random entropy, for which the peak at $S = 6$ implies $2^6 \approx 64$ locations.

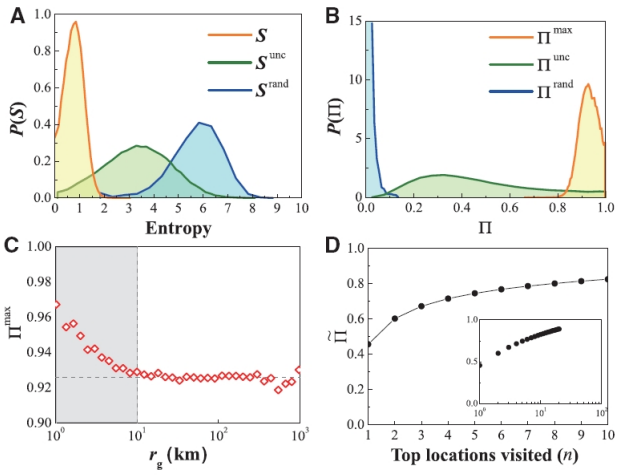

Figure 1.6 **(A)** The distribution of the entropies S, S^{rand} and S^{unc} across 45,000 users. **(B)** The distribution of Π^{max}, Π^{rand} and Π^{unc} across all users. **(C)**. The dependence of Π^{max} on the user's radius of gyration r_g. For $r_g > 10$ km, Π^{max} is largely independent of r_g. **(D)**. The fraction of time a user spends in the top n most visited locations, the resulting measure $\tilde{\Pi}$ representing an upper bound of predictability Π^{max}. Figure from Song et al (2009).

To represent the fundamental limit for each individual's predictability, Song *et al.* defined the probability Π that an appropriate algorithm can predict correctly the user's future whereabouts. If a user with entropy S moves between N locations, then his predictability is bounded by the maximal predictability $\Pi^{max}(S, N)$. For a user with $\Pi^{max} = 0.2$, this means that, no matter how good the predictive algorithm is, only in the 20% of the time can we hope to predict his whereabouts. They determined Π^{max} separately for each user and found that the distribution $P(\Pi^{max})$ is peaked around $\Pi^{max} \approx 0.93$. Figure 1.6**B** highlights that Π^{rand} and Π^{unc} are instead ineffective predictive tools.

Despite the apparent randomness of the individual's trajectories, in a historical record of the daily mobility pattern of the users there is a potential 93% average predictability in user mobility, an exceptionally high value rooted in the inherent regularity of human behavior. The most surprising is the lack of variability in predictability across the popula-

tion, obtained by explored impact of home, language groups, population density and rural versus urban environment. Although the population has an inherent heterogeneity, the maximal predictability Π^{max} varies very little, there are no users whose predictability would be under 80%.

Knowing the history of a person's movements, advanced pattern mining techniques described in chapters 6 and 7 can be used to find patterns and regularities in human mobility, and to foresee his current location with extremely high success probability.

1.2 Social networks and human mobility

In the previous section we presented the evolution of the study on human mobility, describing the main patterns and models that characterize the mobility behavior of individuals. Here, we make a step further in our journey of understanding human behavior by focusing on the interplay between human mobility and social networks, with the purpose of highlighting to what extent human movements affect social dynamics, and how social interactions influence the way people move.

We will first present a brief overview of network science and its growth in the last decade, and then we will focus on the recent developments and discoveries regarding the interplay between the social world and the mobility of people.

1.2.1 Introduction to network science

Network science is a truly interdisciplinary field that examines the interconnections among diverse physical, engineered, information, biological, cognitive, semantic and social systems. In mathematical terms a network is represented by a graph $G = \{V, E\}$, where V is a set of n nodes and E is a set of edges that connects V. According to the definition, any system of interacting elements can be represented as a network. The thinking of complex networks was traditionally dominated by random graph theory, first proposed by Erdös and Rényi back to 1950s. The random graph model presented a simple realization of a network: we start with N disconnected nodes, and randomly connect every pair of nodes with probability p, yielding a graph with $pN(N-1)/2$ edges. As data regarding wiring diagrams of real systems started being collected by computer programs in late 1990s, topological information of real networks became increasingly available, prompting many scientists to ask

a fundamental questions: are real networks, from cell to Internet, truly random? Over the past decade, we have witnessed dramatic advances along this direction, leading to the discovery that despite the intrinsic distinctions in the nature and functionality of the nodes and their interactions, many real world networks follow highly reproducible patterns. There are three most studied properties that characterize a real network:

Average path length measures the average steps it takes for one node to reach another node in the network, also commonly referred as diameter of a network. Although real networks often consist of a large number of nodes, they have a very small diameter, which is most known as the "small world" property or "Six degrees of separation". That is, individuals on the planet were separated by six degrees of social contacts. Despite its simplicity, the random graph model well captures this property, predicting the average path length $d \sim \ln N$, where N is the size of the network.

Clustering represents densely connected cliques in a network, which was formally quantified by Watts and Strogatz. They introduced clustering coefficient C_i for node i, that measures the fraction of neighbors of i are also connected to each other. In random graph model, as links are distributed randomly among the nodes, it predicts $C_i = p$. Yet in almost all real networks, the clustering coefficients are significantly higher than the random graph model prediction. To capture the pervasive clustering phenomena, Watts and Strogatz introduced the small-world model, also known as the WS model: start from a regular network, for instance a ring, in which each node is connected to its k nearest neighbors. Let us redirect links with probability p, moving one end of an edge to a new location chosen uniformly at random from the lattice. When $p = 0$, the network is regular lattice, thus characterized by a very high clustering coefficient but a large average path length. On the other end, when $p = 1$, the network is equivalent to a random graph. As we start to increase p from 0 to 1, the diameter of the network quickly shrinks, while the cluttering coefficients remain roughly the same. Therefore, for a wide range of p, WS model gives rise to networks with both high clustering coefficients and small diameter.

Degree distribution, $P(k)$, measures the probability that a randomly selected node has k edges. The random graph model predicts $P(k)$ follows a Poisson distribution, corresponding to a homogeneous network, where every node has roughly the same degree around $\langle k \rangle$. However, a variety of real networks, spanning from the Internet and WWW to scientific citations and actor collaborations, exhibit the 'scale-free' property,

a highly reproducible pattern not accounted for by either random graph model or WS model. That is, $P(k)$ follows a power law $P(k) \sim k^{-\gamma}$. This result indicates that real networks are rather heterogeneous: most nodes in the network have very low degree, while there are a notable number of nodes with a large number of connections. Think about Yahoo! for the Web, ATP protein for metabolic networks, and Heathrow for air traffic network. To explain the possible origin of the observed scale-free property, Barabási and Albert introduced the scale-free model (or BA model) by viewing the network as a dynamical object that evolves with addition of nodes and links to the system, in strong contrast to the static models that dominated the literature before. Imagine an initial network of a small number of nodes m_0. At each time step we add a new node with m edges that link the node to m different vertices already present in the network. The probability that a new node will be connected to node i depends on the connectivity k_i of that node. After t time steps the model leads to a network with $t + m_0$ nodes and mt edges. This network evolves into a scale-invariant state with the probability that a node has k edges following a power law with exponent $\gamma = 3$.

In addition to the measures listed above, the concept of **tie strength** has attracted particular attention in the study of social networks. It was introduced by sociologist Mark Granovetter in 1973 as a "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal service which characterize the tie". He proposed a model of society consisting of small and fully connected circles of friends, linked by strong ties. Weak ties connect the members of these intimate circles to their acquaintances, who have strong ties to their own friends. Since weak ties act as bridges between separate "social microworlds", they play a crucial role in any number of social activities, such as the spreading of information, ideas and diseases, or in finding a job. Conversely, strong ties link persons in intimate and tight communities, affecting emotional and economic support.

The existence of a local coupling between tie strengths and network topology is confirmed by recent research, which exploit the huge quantity of human interactions recorded by modern tools and technologies. A study conducted by Onnela *et al.* analyzed a huge dataset that stores the mobile phone interaction of millions of individuals in a time period of 18 weeks. The researchers inferred a social network from data connecting two users with a link if there had been at least one reciprocated pair of phone calls between them, and defining the strength of a tie as the aggregated duration of calls. Consistent with the Granovetter's

hypothesis, the majority of the strong ties were found within highly connected communities, indicating that users tend to talk for most of their time with the members of their immediate circle of friends. In contrast, most links connecting different communities were weaker than the links within the communities. Moreover, as a consequence of the topological structure of the network, removing the weakest links leads to a rapid network's sudden disintegration, while removing first the strongest ties shrinks the network but will not precipitously break it apart.

The interesting findings discovered by the above cited study, together with that of more recent works, confirm the importance of tie strength in study of networks, suggesting that weak and strong ties play a different but crucial role in the understanding of many dynamic processes regarding our society.

1.2.2 Interplay between human mobility and social networks

Recent advances on human mobility and social networks have turned the interplay between these two aspects into a crucial missing chapter in our understanding of human behavior. To make progress along this direction requires large-scale data that simultaneously capture the dynamical information on individual movements and social interactions. Thanks to the increasing availability of Mobile phone datasets and location-based online social networks (LBSN, see also Chapter 16), scientists start to look into the questions of to what extent human mobility patterns shape and impact our social ties, and how do our social surroundings affect where we go? The central hypothesis here is that social interactions increase with physical proximity. Indeed, social links are often driven by spatial proximity, from job- and family-imposed shared programs to joint involvement in various social activities. These shared social foci and face-to-face interactions, represented as overlap in individuals trajectories, are expected to have significant impact on the structure of social networks. There are three lines of inquiry in current literature: (1) geographic propinquity yields higher probability of forming a tie; (2) overlap in trajectories predicts tie formation; (3) Social environment affects individual mobility.

Geographic propinquity

The considerable influence of the geographic distance on the formation, the evolution and the strength of friendships is probably rooted in the

very nature of our social brain. According to the anthropologist Robin Dunbar, there is a physical cognitive limit in the number of strong ties our brain is able to manage, partly because they must be powered by a form of social grooming, a time-consuming activity mainly based on geographical proximity and face-to-face contacts.

Recent analysis on Facebook and email data confirmed Dunbar's intuition, showing that the volume of communications is inversely proportional to geographic distance and that the probability $P(d)$ of having a friend at a certain distance decrease following a sort of "gravitational law". Although in the last decades technology has contributed to reduce distances, proximity is still important for the establishment of relevant relationships, breaking down the illusion of living in "a global village": a small world in which physical and cultural distances vanish and where lifestyle become homogeneous.

In studying the social vs geography problem, data from LBSNs proved to be very useful. Scellato et al. used information from both the social and location components of several LBSNs to identify the relation between friendship and geographic distance. They noticed a weak positive correlation between the number of friends and their average distance, and observed that the socio-spatial structure of the users can not be explained by taking into account separately geographic factors and social mechanisms. Cranshaw et al. studied the entropy related to LBSNs locations in order to understand how it affect the underlying social network. They found that co-locations at high entropy locations are much more likely to be random occurrences than co-locations at low entropy locations. So, if two users are only observed together at a locations of high entropy (for example a shopping mall or a university), they are less likely to actually have a link in the underlying social network than if they are observed in a place of low entropy. Moreover, users who visit locations of higher entropy tend to be more social, having more ties in the social network than users who visit less diverse locations.

Trajectory overlap

Given that two persons have been on multiple occasions in the same geographic place at the same time, how likely are they to know each other? This is another interesting and open problem about the interplay between sociality and mobility, regarding to which extent social ties between people can be inferred from co-occurrence in time and space.

Crandall et al. studied such problem by analyzing a huge dataset from the popular photo sharing site Flickr, reaching interesting and striking

conclusions. They inferred a spatiotemporal co-occurrence between two Flickr users if they both took photos at approximately the same place and at approximately the same time. Rather surprisingly, they found that even a very small number of co-occurrences can lead to orders-of-magnitude greater probabilities of a social tie. Indeed two users have nearly 5,000 times the baseline probability of having a social tie on Flickr when they have just five co-occurrences in a day in a 80 km range of distance. With the aim of a deeper understanding of the underlying phenomenon, they developed a mathematical model in which the probabilities of friendship as a function of co-occurrence qualitatively approximate the distributions they observed in the Flickr data.

Wang *et al.* presented a data mining approach to the question of to what extent individual mobility patterns shape and impact the social network. Following the trajectories and communication patterns of approximately 6 Million mobile phone users over three months, they defined three group of similarity measures: mobile-homophily (similarity in trajectories), network proximity (distance in the call graph) and tie strength (number of calls between two users). Exploring the correlation between these measures, researchers discovered that they strongly correlate with each other. The more similar two user's mobility patterns are, the higher is the chance that they have close proximity in the social network, as well as the higher is the intensity of their interactions. Starting from these results, they designed a link prediction experiment, constructing the entire repertoire of both supervised and unsupervised classifiers, based either on network and/or mobility quantities. Results showed that mobility on their own carry high predictive power, comparable to that of network proximity measures. By combining both mobility and network measures, in the supervised case authors obtained that only approximately one fourth of the predicted new links were false positives, and only one third of the actual links were missed by the predictor.

The results of the study by Wang *et al.* suggest that Granovetter's theory should be integrated with a "mobility" dimension: as we can notice in Figure 1.7 the strength of a tie is correlated not only to social proximity (the extent to which people share the same community) but also to their mobility behavior (the overlapping of their spatiotemporal trajectories).

Social environment affects individual mobility

Leskovec et al. investigated the interaction of the person's social network structure and their mobility using datasets that capture human move-

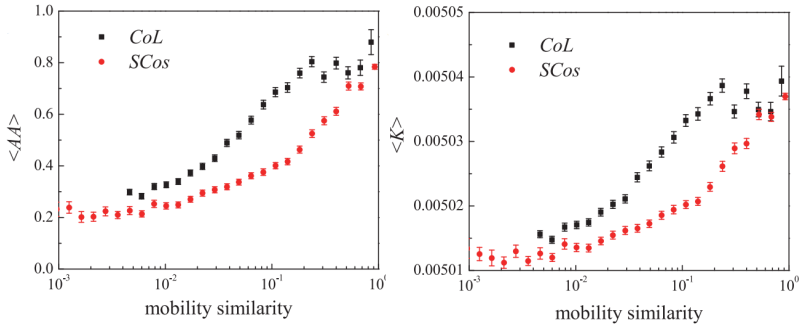

Figure 1.7 Correlations between mobility measures and Adamic-Adar coefficient (left), tie strength (right). The proximity measure used are the spatial co-location (CoL) and the spatiotemporal co-location ($SCos$) inferred from the trajectories of the users. Figure from Wang et al (2011).

ments from Gowalla, Brightkite and phone location trace data. Since they uncovered a surprising increase of the effect of distant friends on an individual's mobility, they tried to understand if friendships influence where people travel, or if it is more traveling that influences and shapes social networks. In order to measure the degree of causality in each direction, they downloaded the Gowalla social network at two different time points t_1 and t_2, three months apart. Considering friendships at time t_1, they calculated a set of checkins C_a that occurred after time t_1 and quantified the influence of sociality on future movements by measuring what fraction of them occurred within the vicinity of friend's homes. Similarly, researchers examined the influence of mobility on creating new social ties by examining a set of checkins C_b before time t_1 and counted the fractions of checkins led to creation of new friendships. They found that whereas there is, on average, a 61% probability that a user will visit a home of an existing friend, the probability that a checkin will lead to a new friendship is only 24%. Such results were confirmed in phone call data, with the influence of friendship on individual's mobility about 2.5 times greater than the influence of mobility on creating friendships. Moreover, data also display a strong dependency between probability of friendship and trajectory similarity, suggesting the there is a strong presence of social and geographical homophily.

The most interesting aspect of such main findings in the interplay between sociality and mobility, is that they can be used to develop a model of human mobility dynamics combining the periodic daily movement

patterns with the social movement effects coming from the friendship network.

1.3 Data mining and network science: a vision of convergence

We have discussed in this chapter how the tools of statistical physics and complexity science have been applied to the study of human mobility, both focusing on individual movements and considering also the social relations among individuals. We have observed how, in both cases, general laws can be devised and empirically validated based on the newly available mobility data, shedding a new light on the underlying mechanisms behind phenomena that, at first sight, seem to be governed by chaos.

We conclude with an observation that spontaneously emerges from the current trend of research, as presented here: there is an evident push towards the convergence of network/complexity science and data mining research, a progressive merge of the two scientific communities that is only beginning today, but it is steadily increasing due to the advantages of combining the complementary strengths and weaknesses of the two approaches. Why this merge is convenient?

We learned in this chapter that statistical physics and network science are aimed at discovering the global models of complex social phenomena, by means of statistical macro-laws governing basic quantities; the ubiquitous presence of power laws and other long tailed distributions witness the behavioral diversity in society at large, such as the huge variability and individual differences of human movements. On the other hand, data mining is aimed at discovering local patterns of complex social phenomena, by means of micro-laws governing behavioral similarity or regularities in sub-populations, such as the mobility patterns and clusters discussed in Chapters 6 and 7 of this book. This dualistic approach is illustrated in Figure 1.8. In the overall set of individual trajectories across a large city we observe a huge diversity: while most travels are short, a small but significant fragment of travels are extraordinarily long; therefore, we observe a long-tailed, scale-free distribution of quantities such as the travel length and the users' radius of gyration. Despite this complexity represented in the data, mobility data mining can automatically discover travel patterns corresponding to set of travelers with similar mobility: in such sub-populations the global diversity vanishes

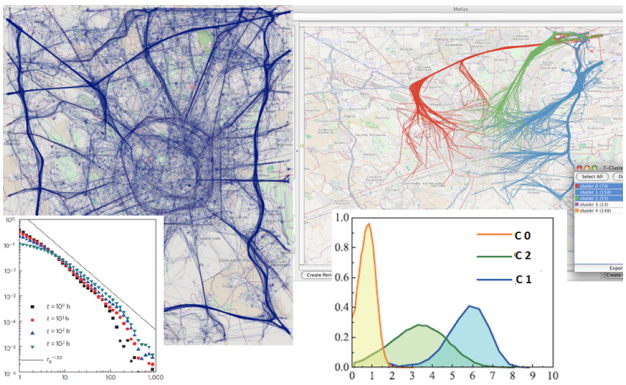

Figure 1.8 The GPS trajectories of tens of thousand cars observed for one week in the city of Milan, Italy, and the power-law distribution of users radius of gyration and travel length (left); the work-home commuting patterns mined from the previous dataset by trajectory clustering and the normal distribution of travel length within each discovered pattern (right).

and similar behavior emerges. The above dual scenario of global diversity (whose manifestation is the emergence of scale-free distributions) and local regularity (within clusters, or behavioral profiles) is perceived today as the signature of social phenomena, and seems to represent a foundational tenet of computational social sciences. Although network science and data mining emerged from different scientific communities using largely different tools, we need to reconcile the *macro/global* approach of the first with the *micro/local* approach of the second within a unifying theoretical framework, because each can benefit from the other and together have the potential to support realistic and accurate models for simulation and what-if reasoning of social phenomena. This vision of convergence among computer science, complexity science and the social sciences is shared today by large research initiatives, such as the FuturICT program[b].

[b] http://www.futurict.eu

1.4 Bibliographic notes

Erdös and Rényi (1959) is the seminal paper that introduced random graphs. The famous small-world model was presented in Watts and Strogatz (1998), while the first argumentations on the small-world phenomenon and the cliquishness nature of society, can be found respectively in Milgram (1967) and Granovetter (1973). The scale-free model was introduced firstly in Barabási and Albert (1999).

The analysis of human mobility based on dollar movements can be found in Brockmann et al (2006). In González et al (2008), are described the mobility patterns discovered by analyzing a rich mobile phone dataset, a work later extended in Song et al (2010). Limits on predictability of human mobility are presented in Song et al (2009), while Karamshuk et al (2011) classifies mobility patterns in temporal, social and spatial dimensions. Cranshaw et al (2010) studies the entropy related to LBSN locations in order to understand how it affect the underlying social network. Crandall et al (2010) analyzed a dataset from Flickr and discovered that even a small number of co-occurrences lead to high probability of a social tie. Wang et al (2011) presents a data mining approach to the question of to what extent individual mobility patterns shape and impact the social network. In Leskovec et al (2011), authors investigate the interactions between social network and mobility by analyzing datasets from location based social network and a mobile phone network.

References

P. Erdös and A. Rényi, *On random graphs*, i. Publicationes Mathematicae (Debrecen), 6:290-297, 1959.

D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature 393, 440 (1998).

S. Milgram, *The small world problem*, Psychol. Today 2, 6067 (1967).

M. S. Granovetter, *The strength of weak ties*, America Journal of Sociology, Volume 78, Issue 6 (May, 1973), 1360-1380.

A. L. Barabási, R. Albert, *Emergence of Scaling in Random Networks*, Science 286, 509 (1999).

D. Brockmann, L. Hufnagel, T. Geisel, *The scaling laws of human travel*, Nature 439, 462465 (2006).

M. C. González, C. A. Hidalgo, A. L. Barabási, *Understanding human mobility patterns*, Nature 454, 779-782 (2008).

C. Song, T. Koren, P. Wang, A. L. Barabási, *Modelling the scaling properties of human mobility*, Nature Physics, 2010

C. Song, Z. Qu, N. Blumm, A. L. Barabási, *Limits of Predictability in Human Mobility*, Science 327, 1018-1021 (2009)

D. Karamshuk, C. Boldrini, M. Conti, A. Passarella, *Human mobility models for opportunistic networks*, IEEE Communication Magazine, May 2011

J. Cranshaw, E. Toch, J. Hong, A. Kittur, N. Sadeh, *Bridging the gap between physical location and online social networks*, In Proceedings of the 12th ACM international conference on Ubiquitous computing, ACM, New York, NY, USA, 119-128, 2010

D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, J. Kleinberg, *Inferring social ties from geographic coincidences*, Proceedings of the National Academy of Sciences, Vol. 107, No. 52 (8 December 2010), pp. 22436-22441.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, A. L. Barabási, *Human Mobility, Social Ties, and Link Prediction*, KDD 2011.

E. Cho, S. A. Myers, J. Leskovec, *Friendship and Mobility: user movement in location-based social networks*, KDD 2011.