Michael Szell, Benedikt Groß

Hubcab – Exploring the Benefits of Shared Taxi Services

Using Big Data to Research Human Mobility

Human activity data has a huge potential for improving urban traffic systems. Data from human interactions such as phone calls, credit card transactions, or social networks, reveals regularities, high predictability (Song et al. 2002), and uncovers the hidden statistical "laws" behind our everyday behavior and movement patterns. The study of human mobility today is no longer limited to the social sciences; researchers from computer science, mathematics, or physics study our movement patterns as they would elementary particles of inanimate matter. Surprisingly, the aggregate and statistical behavior of highly complicated organisms such as humans can in many situations be understood and formulated in purely mechanistic terms (Ball 2003).

But how come all these terabytes of human mobility data now so easily available, and what are the consequences for designing transportation systems? This immediate availability is rooted in both technical and social advances: mobile phones have revolutionized the way in which we communicate: we are always available – and tracked. Telecommunication providers store the position of the nearest cell tower during every call we place or for every SMS we send. Global Positioning System (GPS) technology, equipped on every modern phone and mobile electronic device, allows an even more precise tracking of positions. Not only the paths of walking individuals can be tracked, but paths of individuals using any form of urban transportation. Installed on vehicles, trackers can paint powerful pictures of the flows of millions of vehicles within a city (figure 1) and increase our understanding of collective human behavior and the bottlenecks of our transportation systems.

On the side of social developments, let us highlight the so-called open data or open government initiatives. Open data refers to the concept that certain data should be freely available to everyone to use as they wish, without restrictions from copyright, patents, or other mechanisms of control. Several administrative and governmental institutions have demonstrated in a number of recent cases that making data sets public can be a smart move: examples include open data from the public transportation system of cities, which have spurred the independent development of smartphone apps for live bus and train schedules; the Open311 initiative allowing citizens to more directly interact with their cities;[1] or various repositories aimed at use for scientists in various fields as listed in the Open Access Directory.[2] In a similar spirit, OpenStreetMap is a collaborative project mapping and publishing cartographic data under an open license.[3] In the United States, the modern idea of open government goes back to the year 1966, when a federal law known as the Freedom of Information Act (FOIA) was signed by President Lyndon B. Johnson, following ten years of congressional hearings championed by Congressman John E. Moss to gain access to deliberations of the executive branch under the Eisenhower administration (Blanton 2002). In his struggles, Moss was backed by high-profile journalists seeking to eliminate the bureaucratic hurdles and governmental secrecy of the time – as Moss remarked poignantly, “You had a hell of a time getting any information” (Kennedy 1978). The FOIA allows for the disclosure of previously unreleased documents controlled by the United States government and defines mandatory disclosure procedures. While it has been amended many times since, the main principle is still in effect today, allowing studies such as ours. Watergate and other far-reaching scandals have ensured the continued existence of FOIA and remain a catalyst for open information movements worldwide (Blanton 2002).

[1] http://www.open311.org

[2] http://oad.simmons.edu/oadwiki/Data_repositories

[3] http://www.openstreetmap.org

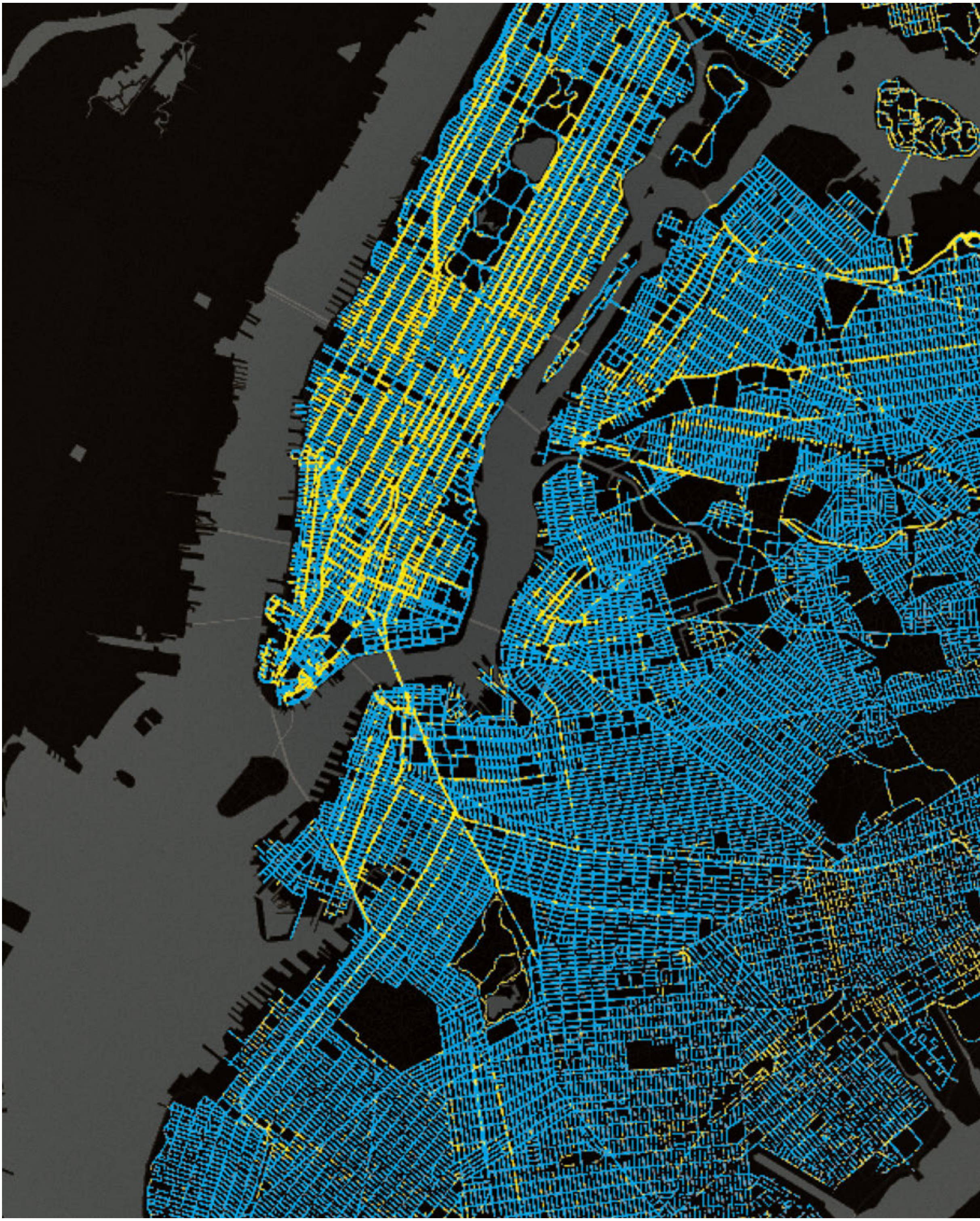
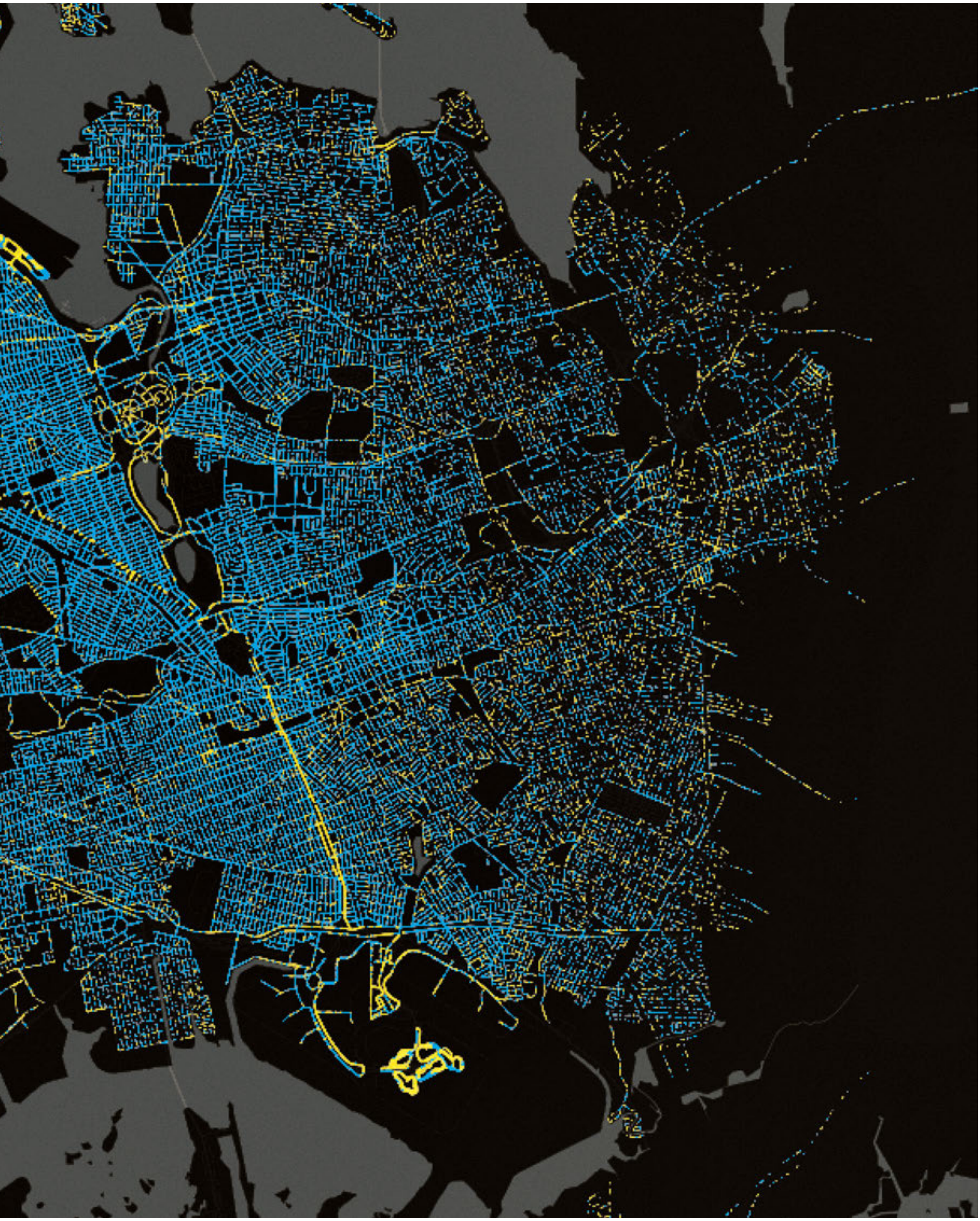


Fig. 1.
Taxi traces in New York City, screenshot from Hubcab. The streets of New York City are partitioned into 40 m segments, all taxi activity of the city over the full year 2011 is visualized.
A segment is shown in yellow if it contains more pickups, in blue if it contains more drop-offs.



The thickness of street segments is proportional to the taxi activity; almost all roads are visited by taxis. Arterial roads tend to emerge as yellow while minor roads are blue, painting a beautiful picture of a complex urban system.

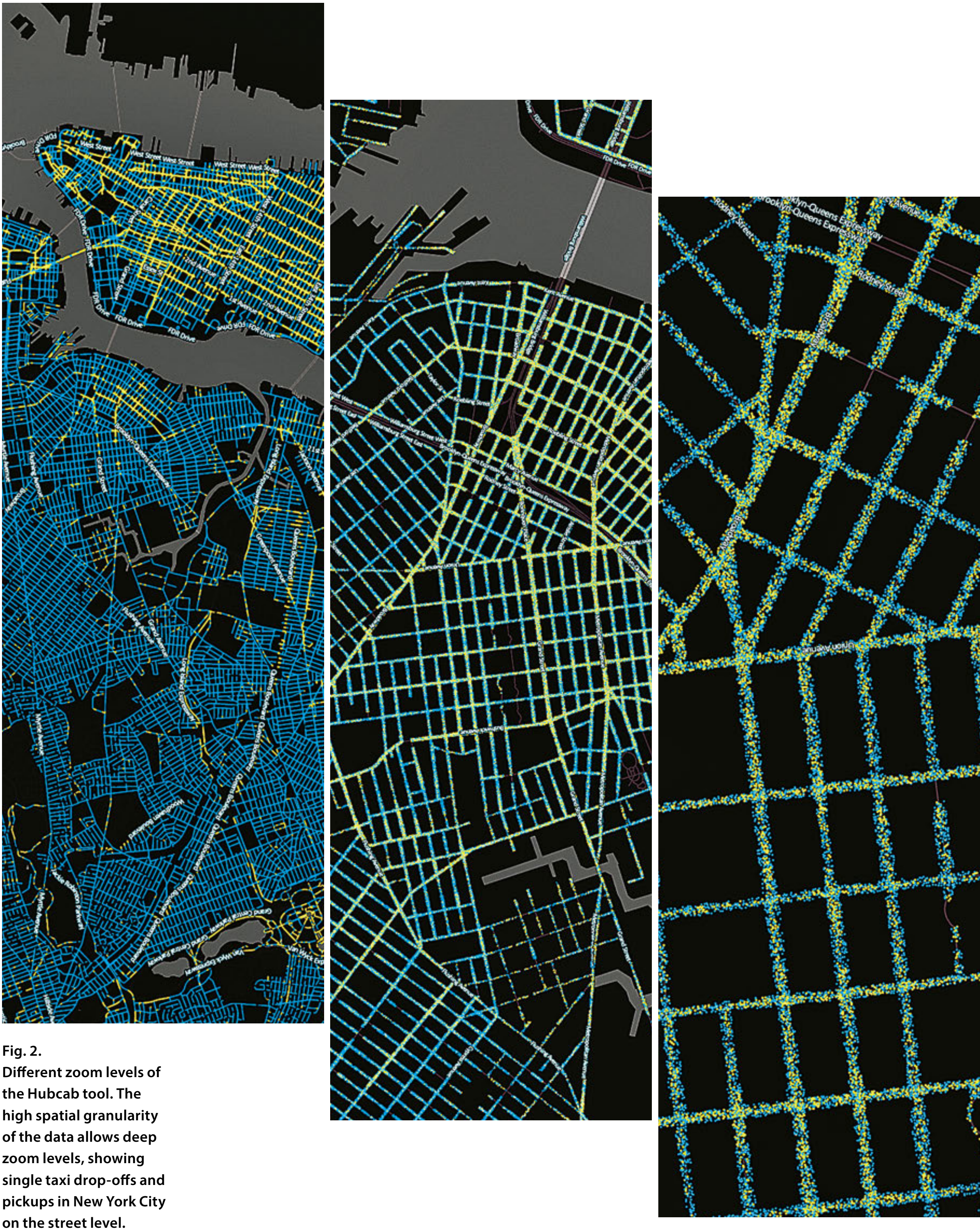


Fig. 2.
Different zoom levels of the Hubcab tool. The high spatial granularity of the data allows deep zoom levels, showing single taxi drop-offs and pickups in New York City on the street level.

Besides such desirable systematic initiatives for more available data in the public interest, data sources for researchers are still often available only through private or unplanned disclosures. For example, mobile phone information is usually proprietary data owned by telephone companies – research teams to persuade companies that data of public interest be disclosed for research purposes. A prime example for an unplanned disclosure is the Enron Corpus, a large database of over 600,000 e-mails written by 158 employees of the Enron Corporation, made public after the investigation of the company's scandal-ridden collapse in 2003.[4] Coordinated social network analysis of the corpus had uncovered previously undocumented corporate practices and gave insight into the nature of group dynamics in large organizations in general. In the context of mobility, the case of Apple iPhones' tracking and storing easily accessible logs of their owners' movements over several months, discovered in April 2011,[5] serves as a prime example for an unplanned disclosure of human movement data. The discovery of these logs led to initiatives such as crowdflow.net to create an open database of Wi-Fi and cell networks.[6]

Improving Urban Taxi Systems

With so much data available, how can we use it to improve life in cities? We are interested in the workings of urban transportation systems. The rapid increase of urbanization and traffic volume in cities makes the smooth functioning of city-based transportation services more vital than ever before. The effect of congestion is a major constraint in urban environments and is felt significantly both economically and environmentally, with emissions impacting regional air quality.

As one of the major modes of urban transportation, let us here focus on urban taxi services. Despite its importance in urban transport, the taxi industry has drawn limited academic interest and has successfully resisted systemic changes and improvements in the past. For example, several New York City mayors had vowed to improve the system over the course of a number of decades but the industry remained essentially the same since the 1900s (Li 2006). However, it is of crucial interest to improve the efficiency of taxi dispatch systems for taxi companies

[4] John Markoff, "Armies of Expensive Lawyers, Replaced by Cheaper Software," *New York Times,* March 5, 2011, A1 (http://www.nytimes.com/2011/03/05/science/05legal.html).

[5] Alasdair Allan, "Got an iPhone or 3G iPad? Apple Is Recording Your Moves," *O'Reilly,* April 20, 2011 (http://radar.oreilly.com/2011/04/apple-location-tracking.html).

[6] Apple updated its operating systems a few weeks later, ending this possibility.

and their clients, as well as urban planners – for example, to decrease fuel consumption using dynamic allocation processes, to improve service quality, and to improve the sustainability of urban traffic.

In a recent research project we studied and made use of a data set containing the positions of all 13,500 taxicabs of New York City (NYC) recorded during all 170 million taxi trips over the year 2011. The taxi business in NYC is regulated by the NYC Taxi and Limousine Commission (TLC) which issues strictly limited permissions (called "medallions") to cabs, granting the familiar yellow vehicles the exclusive right to pick up passengers in response to a street hail. Starting in 2008, all NYC taxicabs have been mandated by the TLC to be equipped with a so-called Taxi Technology System (TTS) including a GPS tracker with a live map of the vehicle's location shown to the passenger on a screen on the backseat. The GPS data collected by the trackers is submitted directly to the TLC. Installation of the GPS trackers was not without controversy – occasional strikes and lawsuits from taxi drivers against the TLC followed and have been ongoing for years, describing the installations in one class action complaint as "unconstitutional, warrantless invasions of their privacy," after a number of medallion licenses were revoked by the TLC for the alleged overcharging of passengers on the basis of the collected data.[7] Regardless, the TLC succeeded in its mandate and complete, anonymized data sets of cab positions can now be freely requested by anyone from the TLC via FOIA request.

NYC is not the only city where positional data of its taxi fleet is available. In the Singapore-MIT Alliance for Research and Technology (SMART) initiative, the largest taxi provider of the city-state of Singapore has granted selected partners access to data collected from 16,000 out of the total 26,000 Singaporean cabs, for studies aimed to improve quality of life within the city. Similar studies exist with data from Shanghai, San Francisco, or Vienna, and more are expected in the coming years. Due to the high-grained resolution and completeness over one full year, we focused first on data from the taxi fleet of NYC; however, our results shall apply without loss of generality to urban taxi fleets in any arbitrary major urban zone.

[7] See Case 1:12-cv-00784-LAK, *Aka and Carniol v. Yassky et al.*

Analyzing data from NYC immediately reveals that the taxi system as a whole is highly inefficient, as found in other cities as well. Inefficiency can be quantified by the total length or time of empty (i.e. no passengers) trips. Empty trip lengths typically display bimodal distributions: The first smaller peak corresponds to the expected average empty trip; the second peak can be found around the distance of city center to airport. It is these special zones of airports and city centers which play a major part in causing the inequilibria between supply and demand. Taxi drivers tend to accumulate at zones where pickups can almost always be expected, such as major roads or other high-throughput locations like airports. However, due to slow propagation of information, variations in actual pickups can become highly increased – especially at airports, queues of empty taxis can get very long at certain times when many drivers get the same idea and go there in unison. This inefficiency leads to adverse effects such as higher-than-necessary emissions, congestion, bottlenecks, etc. The consequences might lead to a reconsideration in urban planning or a strategically superior placement of these areas in future cities.

Several improvements of taxi systems have been attempted in the past, most often in a self-organized manner. Ride-sharing of taxis is a well-known phenomenon in many Asian cities, where passengers split the costs for giving up some comfort and travel time. More recently, commercial attempts at intermodal transport, i.e., the combination of different transportation options, including taxis, as in Daimler's Moovel,[8] or car-sharing services have been introduced, however often suffering from acceptance or similar efficiency problems. Further, a number of smartphone apps have been developed to easily sense taxis, hail taxis, and pay for taxi trips with the tap of a button. These apps are now being employed in cities worldwide with varying success. Urban taxi systems as a whole do not seem to have shown substantial changes, however.

It is due to this rigidity and inefficiency that we propose a new system, which may replace old ones or work in parallel. For this purpose we use the available data and mathematical modeling, designing the new system to be more efficient, causing fewer emissions, and being more affordable to customers than existing ones. The trips in the NYC data set cover over 99 percent of all the city's

[8] http://www.moovel.com

streets (except for parts such as Staten Island which are not serviced by taxis). Each street segment has a unique set of destination points and origin points (street segments were taxis are going to or coming from). In particular we noticed that a large number of trips could potentially be combined and therefore be saved. In dense areas of Manhattan many pairs of street intersections exist where up to 2 million trips are started and ended in the immediate vicinity of the start and the end point (over one year). A large number of these trips start and end at similar times, making them redundant if the passengers were willing to share a cab. Therefore we focused our approach on developing a new dispatch algorithm with the basic idea of combining trips.

This new process aims to minimize the total cost of the service while satisfying certain comfort criteria, e.g., customers should not have to wait for an additional percentage of time than before. In the simplest case a taxi has a capacity of two passengers. When two passengers are found at the same starting point wishing to be delivered to the same end point, both at the same time, the taxi is able to pick up and deliver both. Some assumptions may of course be relaxed, i.e., the start and end points do not necessarily have to match exactly, or may just lie along the route of the longer trip. In any case, we first show via simulations that the current taxi system may be optimized to solve the demand more efficiently with a cab capacity of two. It is then natural to expand the cab capacity to arbitrary larger values. If models, simulations, and analytical investigations show that a larger capacity is even more efficient, practical implementation of large-capacity cabs may resemble "taxi limousines," where passengers have their own private compartment inside a larger vehicle, similar to buses but with individual, dynamic routes.

Mathematically, our approach defines a trip shareability network where nodes are trips, and links between nodes mean that two trips can potentially be combined following some rules and limitations. When more than two trips are to be shared, the network becomes a much more complex entity with links of higher dimension. The algorithm then solves the so-called maximum matching problem, which leads to an optimal solution for the trip combination task. Exact solutions are computationally not feasible for large graphs, however algorithms exist which are able to approximate optimal solutions in reasonable run time. In this process we define a prolongation parameter that stands for the amount of time up to which a passenger is willing to prolong her trip. If this time is chosen low, quality of service is higher since the passenger has to wait for shorter time, but fewer trips can be combined. The situation is reversed for long prolongation times.

The aim of the study is to assess the potential for shareability without substantially lowering the quality of service. Preliminary results on the set of trips within Manhattan are encouraging: For the case of two trips sharing, if passengers are willing to prolong their trips up to five minutes, more than 90 percent of trips can be shared, reducing total travel distance – thus pollution generated – by more than 40 percent. In the case of sharing three or more trips, the potential for shareability seems even larger. The trade-off of an up-to-five-minutes-longer trip for a halved cost seems a reasonable offer. Note that the trip-sharing algorithm appears straightforward to implement practically in an online dispatch system. After hailing a taxi, a user can receive feedback on her shareability options almost instantly.

While the technical details of a shareable urban taxi system seem feasible to solve algorithmically, achievement of the potentially major political task of convincing both authorities and consumers of its viability remains an open question. The overhead of extending existing technologies could be perceived as high, and incentives for overcoming fears of contacts between alien passengers may have to be created (apart from split costs), such as physically separated, private compartments within each vehicle.

Hubcab – The Tool

Parallel to our research efforts we have developed a complementary online tool, named Hubcab, to give a larger audience the possibility to explore the studied phenomena in an easy and fun-to-use way, available at www.hubcab.org. The tool visualizes the same data set of NYC taxi trips, displaying all 170 million taxi trips and allowing interactive exploration of the city from a fresh perspective, demonstrating the potential of a smart dispatch algorithm to a worldwide audience.

Particular attention was paid to represent the amount and direction of the trips in an intuitive way. Drop-offs and pickups are represented coherently and in a dynamic, hierarchically structured way, revealing the maximum information of single drop-off and pickup points in higher zoom levels. This approach has the advantage that the user can see both the direction and an approximation of the amount of the trips at the same time. For cartographic data, we used maps from openstreetmap.org. A script was developed to subdivide NYC's vast road network into more than 200,000 segments of 40 meters' length to provide a high-resolution experience. The vast majority of these segments contains at least one pickup or drop-off, some up to thousands (over a full year). Hubcab allows users to zoom into the map of drop-offs and pickups, showing single points of pickups and drop-offs in zoom levels of unprecedented magnitude (figure 2). The user can switch between different time slices, showing the full data or data segmented in time highlighting different hot spots of pickup or drop-off zones during different times of the day. The radius of this vicinity can be changed on the fly. All the trips between pickup and drop-off points are selected dynamically; Hubcab allows users to explore and visualize the flows between 40 billion (200,000 × 200,000) pairs of street segments. The Hubcab tool provides a unique insight into the inner workings of the city from the previously invisible perspective of the taxi system with a never-before-seen granularity. Hubcab allows one to investigate exactly how and when taxis pick up or drop off individuals and to identify zones of condensed pickup and drop-off activities. Most important, the visualization of flows between any two points of the city allows citizens to experience the redundancy of a large number of trips and the vast potential for improvement. Hubcab expands and changes the perception of urban space, and potentially the behavior of its inhabitants and visitors. Societal and political implications are apparent; urban planning may significantly change through the use of such tools. The sensed, visualized data can be utilized to better design cities, prototyping new urban futures superior on a systemic level. In our particular case this improvement is expected to lead to less congestion in road traffic, lower running costs, and a less polluted, cleaner environment.

In our approach we have been using methods involving optimization algorithms on graphs. Besides this logistical optimization approach, what other types of solutions to the problem would be possible? Previous research using mobile phone data has shown that the specific space syntax can constitute an important influence on the geography of human activity (Reades et al. 2009), and therefore on the functioning of the transportation system. It might be viable to consider rearranging urban structures for shifting land-use patterns, with the goal of improving urban transportation systems in an integrated approach. In any case, an efficient, systematic improvement will be accomplishable only after measuring the status quo and using rigorous analysis on the obtained data sets.

The authors thank research collaborators Paolo Santi and Giovanni Resta (Institute of Informatics and Telematics of CNR, Pisa, Italy), Steven Strogatz (Dept. of Mathematics, Cornell University, Ithaca, NY), Stanislav Sobolevsky and Carlo Ratti (SENSEable City Lab), the Hubcab Web developers (47nord), as well as the National Science Foundation, the AT&T Foundation, the MIT SMART program, the MIT CCES program, Audi Volkswagen, BBVA, Ericsson, Ferrovial, GE, and all the members of the MIT SENSEable City Lab Consortium for supporting the research.

Ball, Philip. 2003. "The Physical Modelling of Human Social Systems," *ComPlexUs* 1, no. 4: 190–206.

Blanton, Thomas. 2002. "The World's Right to Know," *Foreign Policy* 131 (July/August): 50–58.

Kennedy, George. 1978. "Advocates of Openness: The Freedom of Information Movement." Ph.D. diss., University of Missouri, Columbia.

Li, Sonny Heng. 2006. "Multi-Attribute Taxi Logistics Optimization." Ph.D. diss., Massachusetts Institute of Technology.

Reades, Jonathan, Francesco Calabrese, and Carlo Ratti. 2009. "Eigenplaces: Analysing Cities Using the Space-Time Structure of the Mobile Phone Network," *Environment and Planning B: Planning and Design* 36, no. 5: 824–836.

Song, Chaoming, Qu Zehui, Nicholas Blumm, and Albert-László Barabási. 2010. "Limits of Predictability in Human Mobility," *Science* 327, no. 5968: 1018–1021.