Environment and Planning B: Planning and Design 2012, volume 39, pages 413 – 415

doi:10.1068/b3903ed

Editorial

Visualisation tools for understanding big data [1]

In recent years, fuelled by continuing technological advances, there has been an explosion in the creation and publication of visualisations of what has come to be called 'big data'. In the last editorial (Batty, 2012), we wrote about how the rise of the smart city through its routine instrumentation is leading to a generation of enormous real-time datasets—big data—that have the potential for providing us with entirely new information to reveal the functioning of cities at fine scale and over very short time periods. Understanding such data, however, remains a major challenge. Not so long ago, mapping the distributions of phenomena from such large datasets required weeks of data preparation even before its analysis could begin. Now the volume of data released each day exceeds anything that could be collected in the typical academic lifetime of a generation ago. In fact, informed opinion suggests the world's information (data) is doubling every two years and last year (2011) 1.8 zettabytes was collected, a zettabyte being 2 to the 70th power, or 10 to the 21st power. In fact it is hard to visualise this number, never mind the data. Catone (2011) suggests that it is equivalent to the storage in 57.5 billion 32 GB iPads but, whatever the analogy, the number is too big to comprehend. Needless to say in the face of such proliferation, most of this information will be lost despite the possibility that its long-term storage in digital archives is 'still' theoretically possible.

Private companies that, for the first time, collect more personal information than central government have fuelled this transformation in data production. For example, The McKinsey Global Institute (2011) reports that fifteen out of seventeen business sectors in the United States now hold more data on average *per company* than the Library of Congress. This shift is significant in the context of visualisation as the quality of data, especially with regard to its representativeness, does not necessarily increase with volume and will be a lower priority for many, especially in comparison with national census agencies. This is a major issue for the UK's Office for National Statistics as it seeks to phase out the decennial Population Census. That said, the majority of big datasets are most representative of urban populations and so have much to offer wide-ranging studies of urban complexity. Increasingly, they are also frequently temporal, thus offering the potential for the analysis of flows on an unprecedented scale (Batty and Cheshire, 2011). Data is increasingly about interactions and relations, about networks and connections and it is little surprise that the current cutting edge of visualisation is in visualising networks.

The temporal element of big datasets extends to the increasing number of real-time feeds as cities seek to become smarter and the multiple infrastructures within them become better connected. In London, for example, real-time feeds exist for everything from the current depth of the River Thames through to the position of London Underground trains on the network and waiting times at bus stops. Context can be added to these feeds through timetable information (to help calculate delays), passenger flow information, and a wealth of socioeconomic datasets. In tabulated form these data can easily extend to billions of rows and require hundreds, if not thousands, of gigabytes of storage space. So, in this evolving 'big data' landscape how can we, as researchers, contribute and what does visualisation have to offer?

Although we cannot show it here, but instead direct readers to its display at http://mappinglondon.co.uk/2012/04/17/mapped-every-bus-trip-in-london/, London's bus network is

[1] This editorial was inspired by the panel session at the conference "Smart Cities: Bridging Physical and Digital" conference held in UCL London on 20th April 2012. A report of the meeting is available at http://www.bartlett.ucl.ac.uk/casa/news/2012-04-25-SmartCitiesReview

composed of 114 000 daily trips (single buses completing their route) and these data are directly available from timetables online. Over the course of a year, these trips transport in excess of 2.3 billion passengers (Transport for London, 2011), and it is clear that a better understanding of their dynamics offers the potential for dramatic optimisation of routing, minimising congestion, improving the scheduling of buses, and many related matters some of which are already being implemented by Transport for London. However, the map, in line with a large number of similar visualisations, is purely descriptive—it merely illustrates what is there rather than how it works. It shares much in common with the work of early geographers and explorers whose interests were in the description of often-unknown processes. In this context, the unknown has been the ability to produce a large-scale impression of the dynamics of London's bus network. The pace of exploration is largely determined by technological advancement and handling big data is no different.

However, unlike early geographic research, mere description is no longer a sufficient benchmark to constitute advanced scientific enquiry into the complexities of urban life. This point, perhaps, marks a distinguishing feature between the science of cities and the thousands of rapidly produced big data visualisations and infographics designed for online consumption. We are now in a position to deploy the analytical methods developed since geography's quantitative revolution, which began half a century ago, to large datasets to garner insights into the process. Yet, many of these methods are yet to be harnessed for the latest datasets due to the rapidity and frequency of data releases and the technological limitations that remain in place (especially in the context of network visualisation). That said, the path from description to analysis is clearly marked and, within this framework, visualisation plays an important role in the conceptualisation of the system(s) of interest, thus offering a route into more sophisticated kinds of analysis.

The trips visualised on London's network provide the basis on which to perceive the extent of congestion on the road system at the system's key junctions. When this information is combined with traffic flow data, it provides a real-time basis for exploring how patterns of congestion and routing change and evolve during the working day and over longer time periods. In one sense, this kind of data has been available at crude snapshots in time and at a coarser spatial scales for many years but the fact that we are now able to collect it routinely, almost in real time in some instances and begin to visualise it on the same time cycles, provides us with extremely powerful tools to examine problems that previously have been beyond our ability to even articulate, never mind explore. Currently, we are adding the smart card data on trips made across all types of public transport in Greater London to the timetables data and providing a picture of flows in terms of both vehicles and person movements. In this form the data can be animated to provide the first working models or rather representations of how these flows evolve over many different time scales. As each trip is available with a unique identifier, space–time profiles can be assembled for many millions of travellers and their behaviour visualised. With some seven million passengers (trips) in the system on a typical day, we can generate countless aggregations for the dataset we are currently working with which has data over a six- month period. A brilliant visualisation of the public transport data on the rail system by Jon Reades is available at http://simulacra.blogs.casa.ucl.ac.uk/2012/05/pulse-of-the-city-reboot/ for seven days starting at 4am on a typical Sunday and evolving the flows in ten-minute chunks over the week. With such visualisations, patterns on many spatial and temporal scales can be inferred—clearly the usual peaks during the working week, but entertainment events and such like, as well as the influence of school holidays. From such data it is even possible to examine the behaviour of those with free passes—the elderly and the young—in contrast to more typical travellers of working age.

Finally, there needs to be much more sophisticated visualisation of these kinds of results with respect to error and uncertainty in the data (either due to data quality, as noted above, or model assumptions). Uncertainty is often an important oversight in many of the headline visualisations associated with big data and therefore this offers a further area of contribution from the research community. The ubiquity of data visualisations of social phenomena should be embraced and their popularity harnessed to increase the impact of our work. We do, however, need to see description as the starting point rather than the end point of researching big data and work towards analytical insights through the application of well-trusted methods developed during the 'small data' years.

We now stand at a threshold which has major implications for our science and for the way we plan. Many of our tools in planning and design are constructed to examine problems of cities of a much less immediate nature than the kinds of data that are now literally pouring out from instrumented systems in the city. A sea change in our focus is taking place and, over the last ten years, formal tools to examine much finer spatial scales have been evolving, particularly those dealing with local movement such as pedestrian modelling. But now the focus has changed again for big data is not spatial data per se but like big science, its data relates to temporal sequences. No longer is the snapshot in time the norm. Data that pertain to real time, geocoded to the finest space–time resolution are becoming the new norm and our tools and models need to adapt. Moreover in time, our quest to look at the long-term evolution of cities will be reinvigorated by data from the short term as we begin to look at data not over the minute or the hour but over longer temporal cycles eventually joining up the traditional gold-standard censuses such as those that take place every decade. In fact, it is likely that these longer term snapshots will themselves change as digital data from the short term comes to complement and restructure how we look at data in the long term. But that is another story, for a later editorial, but one that is equally important in our quest to provide an understanding of big data.

James Cheshire, Michael Batty

References

Batty M, 2012, "Editorial. Smart cities, big data" *Environment and Planning B: Planning and Design* **39** 191–193

Batty M, Cheshire J, 2011, "Editorial. Cities as flows, cities of flows" *Environment and Planning B: Planning and Design* **38** 195–196

Catone J, 2011, "How much data will humans create & store this year?" *Mashable Social Media* http://mashable.com/2011/06/28/data-infographic/

The McKinsey Global Institute, 2011 *Big Data: The Next Frontier for Innovation, Competition and Productivity* http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Transport for London, 2011, http://www.tfl.gov.uk/corporate/media/newscentre/archive/20391.aspx