g-root.org: Crowdsourcing Data Collection at Scale using Social Networks, Human Sensing, and Mobile Sensors

White paper submitted to the NSF Workshop on Pervasive Computing at Scale

Niklas Elmqvist
Purdue University
elm@purdue.edu

Background and Experience

I am an assistant professor in the School of Electrical and Computer Engineering at Purdue University since August 2008. Before joining Purdue, I was a postdoctoral researcher at INRIA in Paris, France. I received my Ph.D. from the Department of Computer Science and Engineering at Chalmers University of Technology in Sweden in 2006. My research lies in the intersection of information visualization, visual analytics, and human-computer interaction. My past work has focused on traditional visualization topics—such as multidimensional data, graphs, and interaction techniques—but I have lately become interested in novel technologies and platforms for visualization, including tabletops, reality-based interaction, and pervasive and mobile computing.

Vision

Data has been named the next frontier of computing, and has been likened to the oil of the 21st Century in the sense that it must be refined from crude to be useful. However, despite the terabytes upon terabytes of data that our now-digital world produces every year, my view is that it is not necessarily **scale** that is a problem, but **relevance**. According to the market research institute IDC[1], most of the world's data is replicated (75%), most of it is unstructured (95%), and, therefore, most of it is only of use to a select few individuals. Proper analysis and extraction of this sea of data may cause transformative changes to society, including for environmental awareness, social interaction, and community building, but the hurdles towards these changes are many and major. My vision is to change the game by enabling "grassroot" *social data collection* mechanisms where crowdsourcing and social motivation are used to build and maintain massive and up-to-date repositories of relevant and structured data in real time. Concrete examples of use include mapping Wi-Fi strength on a university campus, collecting the lowest gas prices, or tracking a city's snow or garbage removal performance. Because all data are tied to individuals, typically to friends in your social network, the data is **relevant**, and a publish-subscribe model will enable actors to observe, modify, and even create new channels of real-time data that they are interested in (by type or by distance). The data collection will take one of three forms: (1) using embedded sensors in mobile devices, (2) through fixed sensors embedded in the world (accessed using mobile devices), and (3) using humans as sensors (particularly for collecting intangible or inaccessible measures). As a feedback mechanism for driving the social data collection process, I envision using spatiotemporal visualization methods overlaid on online maps (from Google or Bing Maps), properly adapted for heterogeneous and geotagged data on mobile devices, to provide real-time information about an actor's local context. Finally, we will use text visualization to show comments—such as Twitter feeds—in their geospatial context.

[1] J. F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz. "The expanding digital universe." Technical Report, IDC, March 2008.

Evidence for Major Advances

Crowdsourcing is becoming increasingly used for data collection in a wide range of settings, including for geographical places (CycloPath[2], Gowalla, Facebook Places, etc), collective knowledge (wikis in general and Wikipedia in particular), and graphical design (99designs, crowdSPRING, and GFXContests). Recent developments in the visualization community have focused on social data analysis for the masses, heralded by websites such as Sense.us[3], Tableau Public, and ManyEyes[4]. The step to also providing social data **collection** is not far, and extrapolating from the success of websites such as Wikipedia indicates that this would lead to transformative advances across a wide range of areas such as democracy, environmental awareness, and techno-social interaction. My group has during the fall started preliminary work on building mobile apps for such data collection (using the Android platform) that provide spatiotemporal visualization in real-time of concurrently collected data as an incentive and feedback mechanism.

Details: Crowdsourcing Data Collection

The core component of my vision would be *g-root.org*, a web-based cyberinfrastructure for social data collection. Accessed over the Internet, g-root.org will allow people to create accounts, subscribe to data channels (grouped into a semantic ontology of data types), and start collecting data using the site or a mobile app. All data samples collected will be timestamped and geolocated. Depending on the capabilities of the user's device, the user will be able to measure some metrics automatically We will also install larger and more capable sensors in public spaces in the city of West Lafayette, IN—such as on city buses—that users can connect to and collect readings from using Bluetooth. Finally, to promote collecting data on intangible, subjective, or even affective metrics, we will also enable humans acting as sensors for such data.

Additional examples include attendees of a college football game using g-root.org to "get the pulse" of the other spectators at any point in time using their mobile devices, and participating by giving their own reactions to the game. Pollution measurements from city buses can be appropriated by any bus passenger and collected into the g-root.org site, collectively improving the environmental awareness of the city. Citizen meteorologists can add their own weather data—qualitative and quantitative alike—to improve the reliability of the weather reporting in the local area. Inhabitants of a region in conflict can provide real-time information about locations with unrest.

I anticipate that the following challenges must be explored at some depth:

- **Hardware challenges** – Find sensors for embedding in (or interfacing with) mobile devices as well as embedding in the world (such as on buses and in public spaces, etc).
- **Social impacts and motivation** – Study ways to promote participation through social-psychological incentives, reputation building, and personal relevance of collected data.[5]
- **Privacy and trust** – Correlating data collection with social networking sites such as Facebook to enable access control of data, as well as support anonymization of data.
- **Visualization techniques** – Visual representations for uncertain spatiotemporal quantitative (data values) and qualitative (tags and comments) data adapted for mobile screens.
- **Robustness** – Data collection when internet connectivity is temporarily unavailable.

---

[2] L. Priedhorsky and L. Terveen. "The Computational Geowiki: What, Why, and How". In *Proc. CSCW 2008*.

[3] J. Heer, F. B. Viégas, and M. Wattenberg. "Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization". In *Proc. CHI 2007*.

[4] F. B. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. "ManyEyes: a Site for Visualization at Internet Scale". *IEEE TVCG* (Proc. InfoVis/Vis 2007).

[5] J. Heer and M. Agrawala. "Design Considerations for Collaborative Visual Analytics". In *Proc. VAST 2007*.