3 is Not a Crowd, It's an Anecdote

Jon Crowcroft
Marconi Professor of Communications Systems, Computer Laboratory
University of Cambridge
Cambridge, UK
jon.crowcroft@cl.cam.ac.uk

Abstract

Crowd-sourcing measurements from the Internet, for the Internet, and by the Internet, is naturally popular due to the cost scaling that this approach offers. It appears to offer a low barrier to entry to new researchers, when compared with the difficulties in obtaining data about operational networks from their owners.

In this talk I will outline some precautions that should be taken by researchers, including:

- Sample Bias – Ground truth must be established about your sample. The very fact that you are using people on the internet prepared (assuming you have informed consent – you do, don't you?) to be part of an experiment, means you have selection bias. Establish what it is by old fashioned means.
- Subject Privacy – You may get more buy-in from more users, and in more ways, if you can offer assurances about maintaining their privacy:
  - "No, this app won't (unlike 70% of smart phone apps) leak lots of your Personal Data",
  - How you achieve this, best practice with sandboxes and cryptography and security processes,
  - AAA applied to your logging databases,
  - Long term fate, care and curation.
- Repeatability – as a data scientist, you have a duty to share what data you can, so that others can verify, and build on your work with confidence. As a social scientist, you have a duty not to put users' privacy in double jeopardy.

There is a saying "Two's a company, three is a crowd". The title of my talk reflects the fact that one can construct a graph from a series of basic triangular graphs, but also that one can triangulate nodes in a graph from data associated with other nodes. Re-identification of nodes (e.g. users) in the Internet is the source of many anecdotes concerning loss of privacy. When crowd-sharing data that has been crowd-sourced from the Internet, we should all be extremely aware of this non-trivial risk. In some areas of studies (e.g. censorship, hacktivism,cybercrime), the risk of such re-identification to some members of our crowds may be extremely high. We are all social scientists, now.

C2B(I)D'15, August 17, 2015, London, United Kingdom.
ACM 978-1-4503-3539-3/15/08.
DOI: http://dx.doi.org/10.1145/2787394.2787404

ACM CCS Concepts: C.2, Networks

Keywords: Internet Measurement, Crowdsourcing, Privacy

BIO

Jon Crowcroft is the Marconi Professor of Networked Systems in the Computer Laboratory, of the University of Cambridge. Prior to that he was professor of networked systems at UCL in the Computer Science Department. He has supervised over 45 PhD students and over 150 Masters students.

He is a Fellow of the ACM, a Fellow of the IEEE, a Fellow of the Royal Society, and a Fellow of the Royal Academy of Engineering. He was a member of the IAB 96-02, and went to the first 50 IETF meetings; was general chair for the ACM SIGCOMM 95-99; is recipient of Sigcomm Award in 2009. He is the Principle Investigator in the Computer Lab for the EU User Centric Networking project, the EPSRC funded Hub-of-All-Things IoT project, and is on the Interim Science Board for the new Alan Turing Institute for Data Sciences.

Jon's research interests include Communications, Multimedia and Social Systems, especially Internet related.