Network Analysis on Provenance Graphs

Mark Ebden[1], Trung Dong Huynh[2], Luc Moreau[2], Sarvapali Ramchurn[2], and Stephen Roberts[1]

(1) Department of Engineering Science, University of Oxford

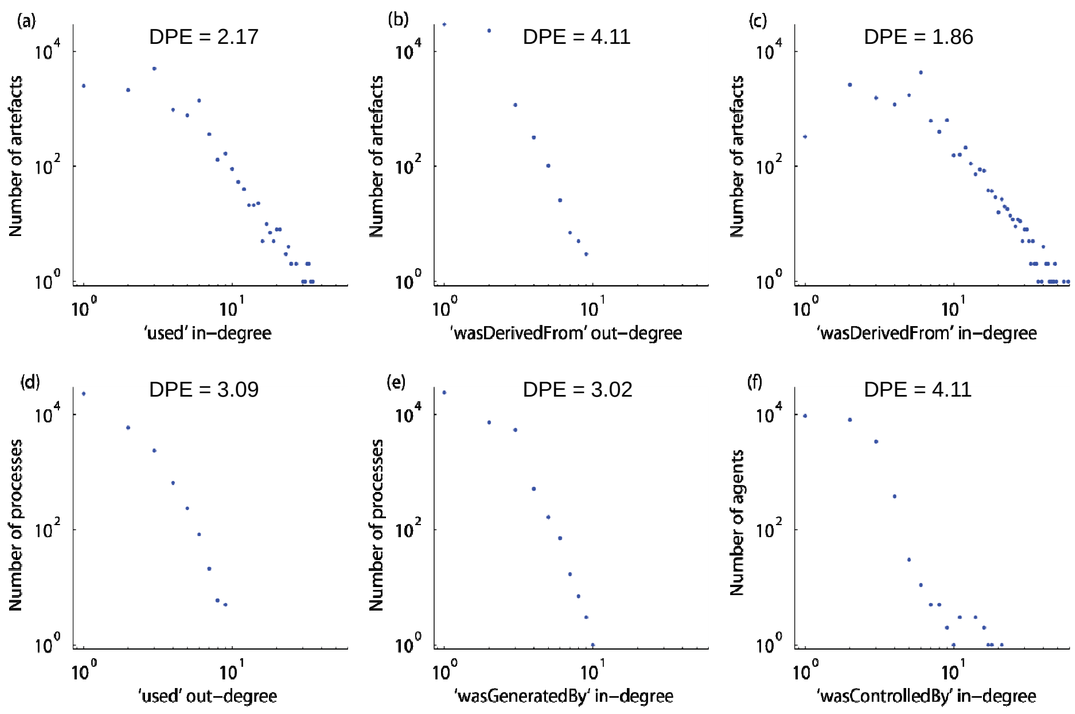
(2) Electronics and Computer Science, University of Southampton

Overview

- Analytical study on various network measures of over 5,000 provenance graphs from the crowdsourcing application CollabMap
- Findings include:
  - Provenance graphs possess similar structural characteristics of real-world networks, such as having small diameters, power-law degree distributions, and a similar densification pattern
  - They are suitable for exploitation of existing network analysis tools for modelling, prediction, and inference
- Provenance-specific network metrics were devised to gain insights about the structure of provenance graphs

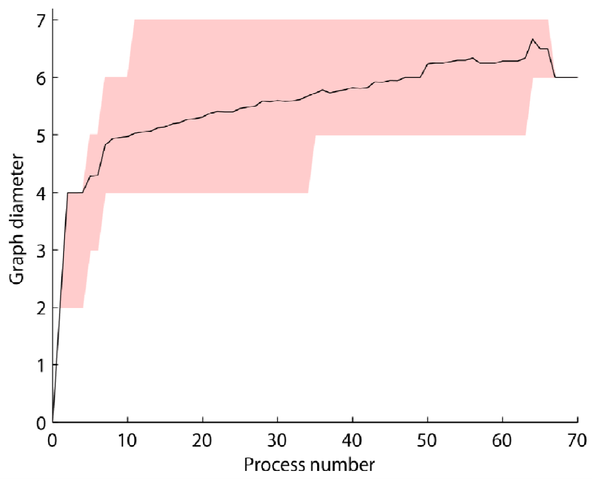
Examples of Practical Features

- Degree-distribution power-law exponent (DPE)



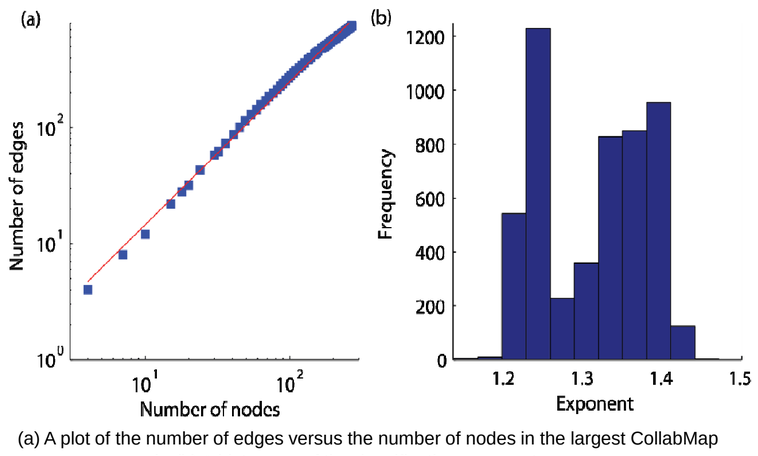
Degree distributions according to edge type, with the values of the degree-distribution power-law exponent (DPE)

- Graph diameter: the longest *distance* in the graph, where the distance between two vertices is the length of the shortest path between them
- Maximum Finite Distance (MFD): the longest shortest path between two node types



Evolution of graph diameter with the number of processes (micro-tasks) for up to 5,128 CollabMap provenance graphs

- Densification exponents: as a network evolves over time (in crowdsourcing, these changes are driven by the sequence of user contributions), it generally becomes denser. The densification power law is $E(t) \propto N(t)^a$
- Edge-to-node correlation (ENC) coefficients between the number of edges and the number of nodes in a growing graph

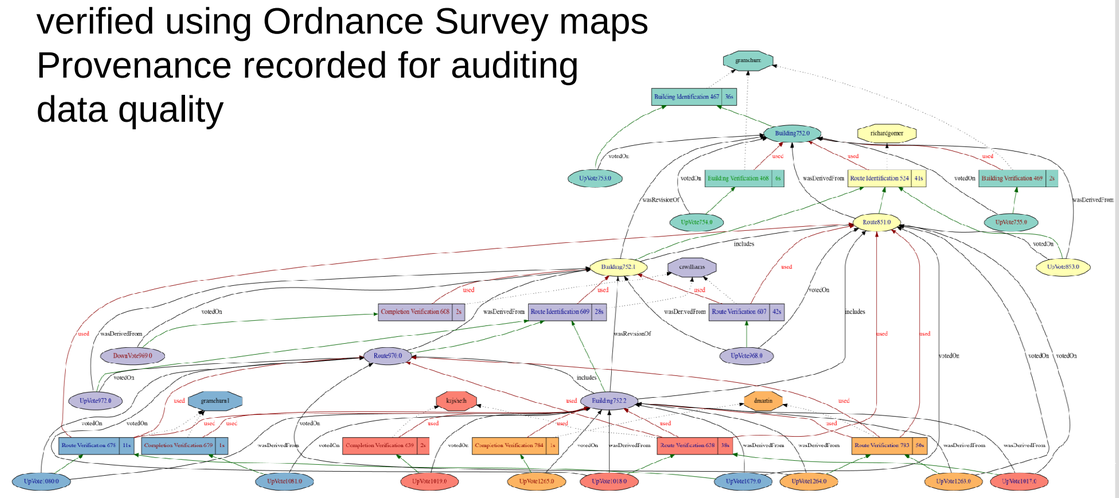


(a) A plot of the number of edges versus the number of nodes in the largest CollabMap provenance graph. (b) A histogram of the densification exponent a.

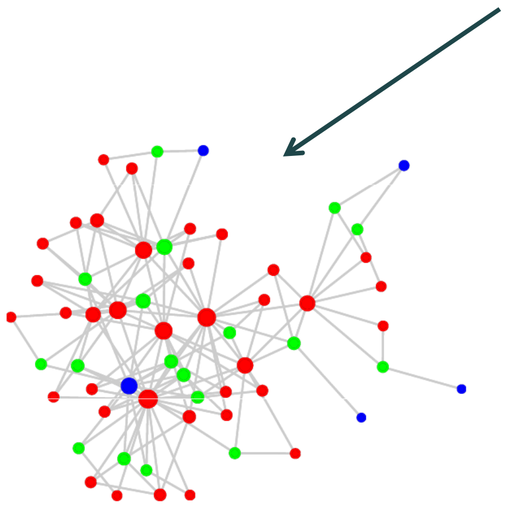
CollabMap Provenance Graphs

Crowdsourcing the identification of buildings and evacuation routes:

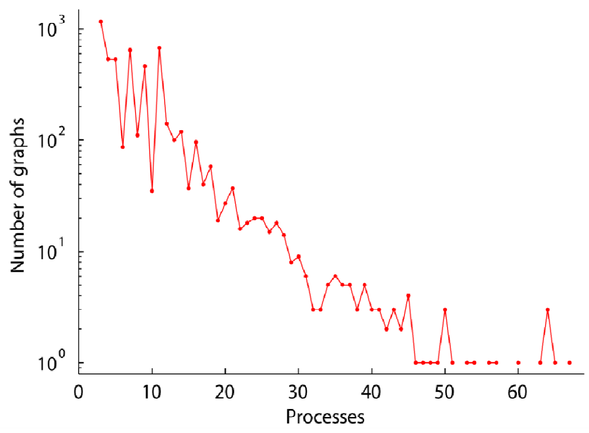
- City-wide mapping of buildings and evacuation routes for disaster-recovery simulations
- Results cross-checked by human users and automatically verified using Ordnance Survey maps
- Provenance recorded for auditing data quality



A provenance graph of a typical CollabMap task in the Open Provenance Model's graphical notations.



A provenance graph, in which artefacts are red, processes are green, and agents are blue.



A plot of the number of CollabMap provenance graphs that contained any given number of processes (micro-tasks).

Current Work

- Classification of provenance graphs, to provide capabilities for capturing, querying, and reasoning over provenance/trust/reputation of the information
- Inference about missing links and nodes, and tracking time-evolving graphs, through these metrics and Kronecker graphs
- Semantic interpretation of the metrics, including their link with agent responsibilities and the degree of inter-dependency among the crowdsourcing activities
- Link to Agile Teaming: e.g. community-detection algorithms on a provenance graph might help to identify collusion among users

References

- Ebden M, Huynh TD, Moreau L, Ramchurn S, Roberts S. Network Analysis on Provenance Graphs from a Crowdsourcing Application. In: *Proceedings of the Fourth International Provenance and Annotation Workshop*, Santa Barbara, USA, 18-22 June 2012.