Network Analysis on Provenance Graphs from a Crowdsourcing Application

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, Oxford, OX1 3PJ, United Kingdom
{mebden,sjrob}@robots.ox.ac.uk
www.robots.ox.ac.uk/~parg

[2] Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom
{tdh,l.moreau,sdr}@ecs.soton.ac.uk
www.ecs.soton.ac.uk

Abstract. Crowdsourcing has become a popular means for quickly achieving various tasks in large quantities. CollabMap is an online mapping application in which we crowdsource the identification of evacuation routes in residential areas to be used for planning large-scale evacuations. So far, approximately 38,000 micro-tasks have been completed by over 100 contributors. In order to assist with data verification, we introduced provenance tracking into the application, and approximately 5,000 provenance graphs have been generated. They have provided us various insights into the typical characteristics of provenance graphs in the crowdsourcing context. In particular, we have estimated probability distribution functions over three selected characteristics of these provenance graphs: the node degree, the graph diameter, and the densification exponent. We describe methods to define these three characteristics across specific combinations of node types and edge types, and present our findings in this paper. Applications of our methods include rapid comparison of one provenance graph versus another, or of one style of provenance database versus another. Our results also indicate that provenance graphs represent a suitable area of exploitation for existing network analysis tools concerned with modelling, prediction, and the inference of missing nodes and edges.

1 Introduction

Crowdsourcing is an increasingly popular approach for tasks that computers find too difficult to solve; the method distributes tasks among human contributors, often through a website. For instance, citizen-science projects at Zooniverse (www.zooniverse.org) have managed to enlist hundreds of thousands of volunteer "citizen scientists" to classify distant galaxies, transcribe historical naval logs, and more. The volunteers contribute data of a quality that is as varied as their backgrounds and expertise. Usually cross-verification among participants helps to discard inaccurate results, yet challenges remain in anticipating how different human contributors will behave and in designing a robust crowdsourcing application.

P. Groth and J. Frew (Eds.): IPAW 2012, LNCS 7525, pp. 168–182, 2012.

CollabMap (www.collabmap.org) is an online mapping application in which we crowdsource the task of identifying residential evacuation routes, with the eventual aim of helping to plan large-scale evacuations in case of disaster. In an effort to address the aforementioned human challenges, we introduced provenance tracking into CollabMap, capturing in detail how contributors trace buildings and draw evacuation routes, and noting the dependencies among their contributions. The resulting provenance graphs allow us to re-create the situations in which the data were generated and to inspect them for potential inaccuracies. In order to gain an understanding of the common characteristics of these graphs, here we carry out an analytical study on various network measures and report our findings. Other researchers have viewed provenance graphs in alternate ways: Altintas *et al.* [1] have analysed them as collaboration networks, and Margo *et al.* [11] have used them as a basis for node classification. The present work offers a deeper level of mathematical abstraction, and our contributions are twofold. First, we estimate probability distribution functions over three selected characteristics of these provenance graphs: the node degree, the graph diameter, and the densification exponent; to our knowledge we are the first to analyse provenance graphs in this way. Second, we devise provenance-specific network measures for provenance graphs, to gauge whether such measures provide a novel insight into provenance graphs, or whether generic network measures are enough. We are also exploring the question of whether provenance graphs, at least those from crowdsourcing contexts, are suitable candidates for existing network methods that support graph modelling, prediction, and the inference of missing nodes and edges.

The remainder of the paper is organized as follows. Section 2 provides a summary of the CollabMap application, including how it works and how provenance was modelled. In Section 3, we describe a range of techniques to extract characteristics from the CollabMap provenance graphs. Section 4 reports our main findings, and the paper is concluded with a discussion in Section 5.

2 CollabMap

In planning the responses to city-wide disaster scenarios, simulating large-scale evacuation is a major challenge, owing in part to the lack of detailed evacuation maps for residential areas. These maps need to contain evacuation routes connecting building exits to the road network, while avoiding physical obstacles such as walls or fences. Existing maps do not provide such routes. To our knowledge, automated techniques to augment current maps with such paths are not available, and direct surveys of city-scale residential areas are usually infeasible owing to the significant effort required. Against this background, CollabMap was developed to crowdsource the drawing of evacuation routes for the public by providing them with two freely available sources of information from Google Maps: aerial imagery and ground-level panoramic views. During a recent two-month trial on the website we established, contributors were awarded cash-prize lottery tickets in proportion to the number of contributions they made. Our ongoing application has so far produced 5,128 provenance graphs for 37,931 micro-tasks completed by over 100 contributors.

2.1 CollabMap Workflow

Based on the Find-Fix-Verify pattern [3], we divide the task of identifying evacuation routes for a single building into smaller activities, called *micro-tasks*, carried out by different contributors. We have designed five types of micro-task:

A. Building Identification. The outline of a building is drawn on the map. It serves as the basis for the other micro-tasks.
B. Building Verification. The building outline is assessed, with a vote of either valid (+1) or invalid (−1).
C. Route Identification. An evacuation route is drawn, to connect an exit of the building to a nearby road.
D. Route Verification. The evacuation route is assessed, with a vote of either valid (+1) or invalid (−1).
E. Completion Verification. The set of evacuation routes is assessed for exhaustiveness, with a vote of either complete (+1) or incomplete (−1).

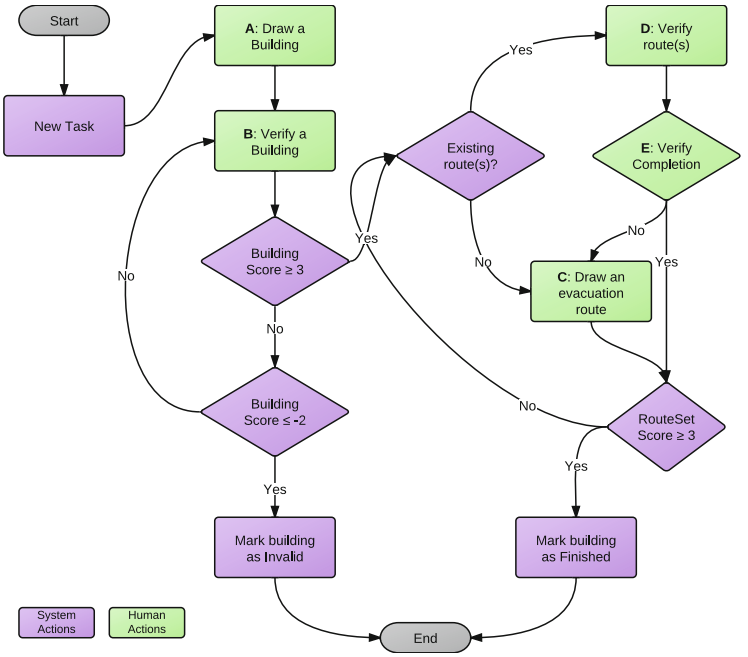

Fig. 1. The CollabMap workflow for identifying evacuation routes of a building

The CollabMap workflow (Figure 1) has two main phases:

Building phase. The outline of a building that has no evacuation route needs to be drawn (**A**). The outline is then checked by other contributors, who vote up or vote down the building outline (**B**) without seeing others' votes. If the total score of the

building, defined as the sum of all the votes, reaches $+3$ then the Building phase ends and the Evacuation route phase begins. If the score reaches -2, the building outline is rejected and marked as invalid.

Evacuation route phase. This is the main activity carried out by CollabMap contributors. The first is permitted only to draw a route (**C**). Subsequent contributors are asked to verify routes (**D**) and are asked whether the set of routes is complete (**E**); if it is not, they are invited to draw new routes (**C**).

In both phases, in order to avoid biases, a contributor is not allowed to verify his or her own work.

2.2 Recording Provenance

We adopted the Open Provenance Model (OPM) [13] for capturing the provenance of data generated in the CollabMap application. The micro-tasks in the previous section generate data of four different types: building outlines, evacuation routes, route sets (collections of routes belonging to a building), and votes. The classes for these data types are **Building Outline**, **Route**, **Route Set**, and **Vote**, respectively (see Figure 2). In order to keep separate the application-specific data from the provenance-related information, OPM constructs were recorded in their own classes: **Artefact** representing a data entity (via the *subject* relation), **Agent** a CollabMap contributor, and **Process** an instance of one of the five types of micro-task above.

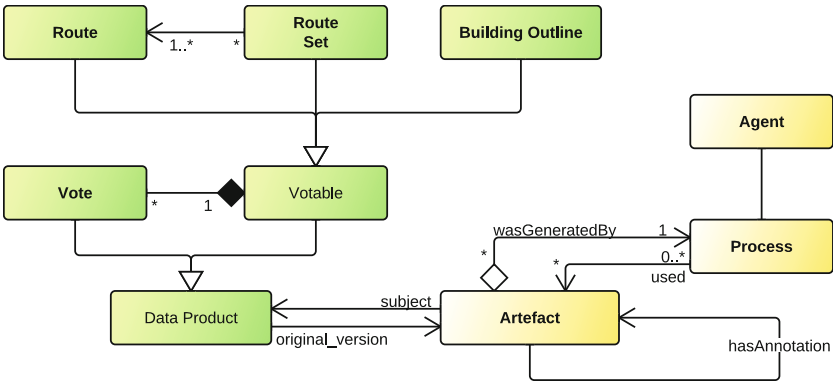

Fig. 2. The UML class model for CollabMap's data and provenance classes. **Data Product** and **Votable** are abstract classes.

When a contributor completes a micro-task, this is recorded as a process along with timing information (namely, how long it takes; see Figure 3 for an example). We also record the artefacts (equivalently, the corresponding data products) that were generated by the process (via the *wasGeneratedBy* relation), and we record which existing artefacts were shown to the contributor in the micro-task (via the *used* relation). Own knowledge of the internal workings of CollabMap also enabled us to assert various

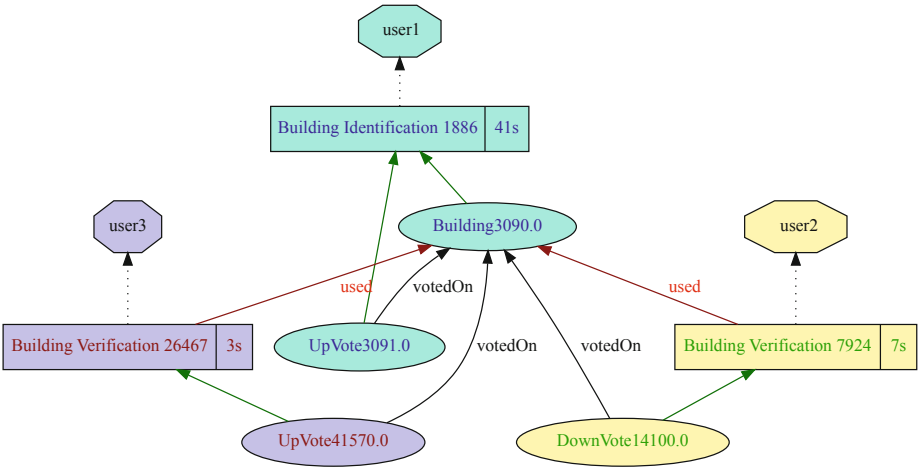

Fig. 3. An example OPM provenance graph recorded by CollabMap showing a building was drawn and voted on by three different users

direct relations between artefacts (via the *hasAnnotation* relation in Figure 2): *wasDerivedFrom*, *wasRevisionOf*, *includes*, and *votedOn*. The last three are special cases of the *wasDerivedFrom* relation, and were treated as such in our analyses in subsequent sections.

2.3 Provenance Graphs

Newman [14] describes four types of network: technological, social, biological, and informational. Provenance graphs fall into the last category, as they are networks describing relationships among elements of information. Other examples of informational networks include those which describe co-authorship of academic articles, semantic relationships among words, and peer-to-peer exchanges of online content. Using the vocabulary associated with the collection of relational network data, our CollabMap provenance graph data are *enumerated* as opposed to being partial or sampled; that is, they are collected in an exhaustive manner from the full population. Our population concerns the totality of the CollabMap data set as of March 2012.

We create a graph $G = (V, E)$, with vertex set V and edge set E. Edges in the present work are unweighted and directed, but our design is extensible to weighted edges, to indicate reliability of connection or other probabilistic phenomena. Five edge types are defined by the OPM: *used*, *wasGeneratedBy*, *wasControlledBy*, *wasDerivedFrom*, and *wasTriggeredBy*. In the current work we recorded all but the last of these, in addition to all three possible node types: artefacts, processes, and agents. Node type is the only vertex attribute currently under study in our provenance graphs, but it is possible to assign additional attributes, either discrete (for example a classification indicating the level of experience of each agent), or continuous (for example, a probabilistic estimate of how often an agent errs during the evaluation of route evacuations).

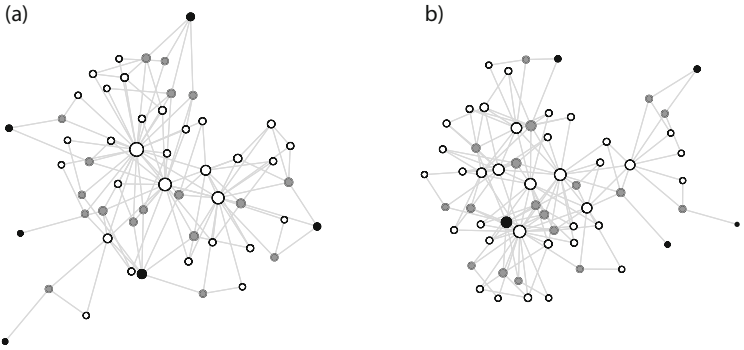

Fig. 4. Provenance graphs for two typical CollabMap tasks, in which artefacts are white, processes are grey, and agents are black. Vertex size increases with degree logarithmically.

To aid graph visualization in Figure 4, vertices in V are represented as circles coloured by node type, and edges in E are represented as straight lines. The graph is drawn in two-dimensional space, but it is possible to imagine the same information appearing in three-dimensional space or on another surface. Vertices are situated according to the Kamada-Kawai free-energy technique in Pajek software [2]. Vertex size is proportional to $\log(d + 3)$, where d is node degree.

The graph in Figure 4(a) contains 54 vertices after 18 processes occurred (18 microtasks), while that in Figure 4(b) contains 59 vertices after the same number of processes occurred. The maximum number of processes occurring in a given provenance graph was 70. Figure 5 gives the distribution of provenance graphs over their maximum process index; it indicates that the majority of tasks were edited at least seven times, and 288 graphs were edited twenty times or more.

3 Methods

To compare the 5,128 networks with those described in the literature, and to see whether the characteristics ascertained from network analysis might be useful, we selected a subset of network properties to investigate. We chose three properties that have been used elsewhere in the analysis of both real and synthetic graphs [10]. They are as follows:

Degree distribution: For many graphs, the degree distribution follows a 'power law' such that the number of vertices N_d with degree d is given by $N_d \propto d^{-\gamma}$, where $\gamma > 0$ is usually called the power-law exponent. We shall examine the degree distribution of an entire provenance graph, and subdivide this into several distributions based on the four edge types and their directionality. In summarizing the information in such plots, we refer to γ as the degree-distribution power-law exponent (DPE), calculated according the method of Clauset *et al.* [5] concentrating on nodes with high degree.

Diameter: The diameter of a graph is the greatest minimum distance between any two nodes. Most real-world graphs exhibit relatively small diameter (the "small-world"

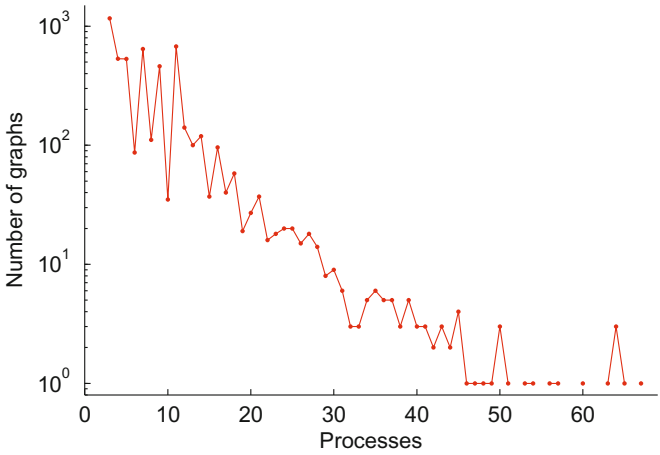

Fig. 5. A plot of the number of CollabMap provenance graphs that contained any given number of processes (micro-tasks)

phenomenon [12]) which tends to stabilize as the number of vertices in a network grows over time (here, as processes occur). Since CollabMap nodes are separated by directed edges, thereby preventing some nodes from forming a path to certain others, strictly speaking the diameter of each graph is infinite; however, by temporarily assuming the edges are undirected, we are able to calculate a diameter and we record its value after each process (micro-task) occurs. In addition, we return to the directed graph to calculate a useful variation on graph diameter: Dijkstra's algorithm [6] provides the minimum path length separating each pair of nodes, and we consider the distribution of the cases in which this path length was a finite number. This distribution determines the maximum finite distance (which we shall refer to as MFD) from one node type to another. We calculate the values of MFD on full provenance graphs as well as on the corresponding data-flow graphs — that is, graphs with only artefacts and *wasDerivedFrom* edges, with no processes involved.

Densification: As a network evolves over time, it generally becomes denser. This can be quantified by comparison of the number of edges to the number of nodes, after each process occurs. The relation between the number of edges $E(t)$ and the number of vertices $N(t)$ in an evolving network after process t ordinarily obeys the densification power law, which states that $E(t) \propto N(t)^a$ for some densification exponent a typically greater than unity [9]. In our provenance graphs, we have chosen to also specialize this relation by node type and by edge type, noting Pearson's product-momentum correlation coefficient in each case. We refer to each coefficient as the edge-to-node correlation (ENC).

Our descriptions of the above three properties have indicated that many graphs which have been studied elsewhere in the literature have a degree distribution following a

power law, have small diameter which stabilizes eventually, and become denser over time in a manner that follows a power law as well. To summarize, our methodology for analysing these three properties on each provenance graph results in several plots and includes the following three metrics: DPE, MFD, and ENC.

4 Results and Discussion

We now present the results from the analyses described in the above section for the provenance graph depicted in Figure 4(a) and for the largest provenance graph. In addition, we carried out the same analyses for the whole population of 5,128 provenance graphs recorded by CollabMap and summarize their results here.

4.1 Degree Distribution

Figures 6(a) and (b) plot degree distributions (histograms depicting how many nodes had a certain number of interconnections) which were typical of those for the provenance graphs under study. The tails (high-degree data) conceivably follow a power law, although the low-degree data points (here, for node degrees fewer than three) lie below this trend; this is a pattern observed in many networks elsewhere [15]. The degree-distribution power-law exponents (DPE) for the data in these two figures were 2.1 and 2.0. Over the 5,128 graphs we examined, the mean DPE was 2.4, with a standard deviation of 0.2. In comparison, elsewhere in the literature values tend to fall between 1.4 [15] and 4.3 [5], with the vast majority between 2 and 3 [15]. The full distribution, given in Figure 6(c), is clearly multi-modal; this is because some of the provenance graphs under investigation were small, and the calculation of DPE is only reliable for large graphs. We found that restricting the analysis to graphs with a minimum size of 40 nodes (recognizing that the maximum number of nodes in a graph was 271) led to the emergence of a peak near DPE = 2.2. In summary, our graphs tended to follow a power law, and the values of DPE were in the typical range.

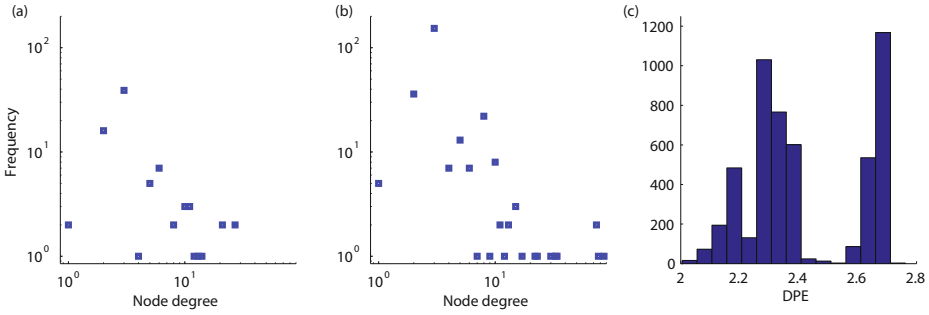

Fig. 6. (a) Distribution of node degrees for the typical provenance graph shown in Figure 4(a). **(b)** A similar distribution for the largest provenance graph. **(c)** Degree-distribution power-law exponent (DPE) for all 5,128 provenance graphs.

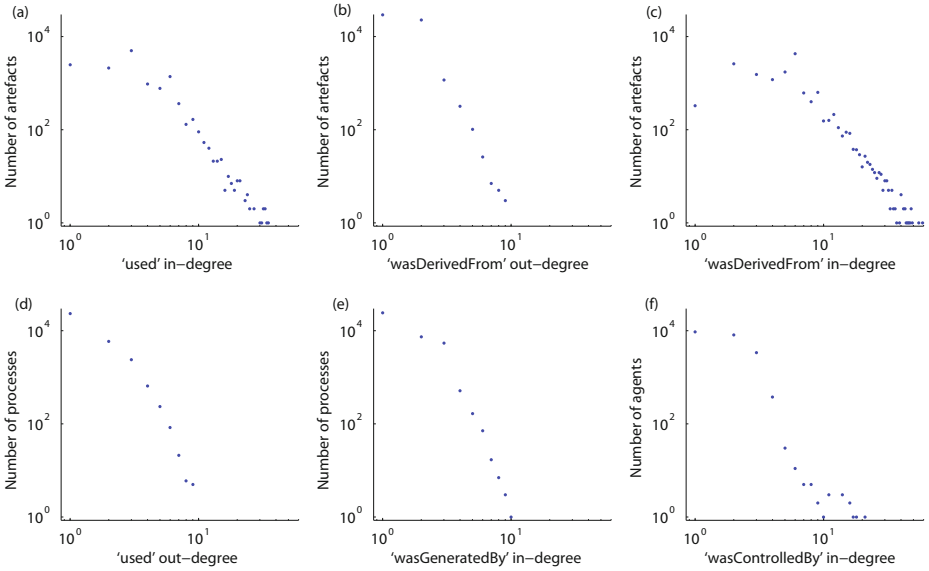

Fig. 7. Degree distributions according to edge type. As the graphs are logarithmic, zeros cannot be plotted; the number of nodes corresponding to zero in-degree or zero out-degree were as follows: (a) 45,194; (b) 5,128; (c) 44,390; (d) 5,157; (e) 34; (f) 0.

Figure 7 shows the degree distribution specialized according to edge type. This figure is probably one of the most useful from a provenance point of view. Since we take into account the directedness of the graph edges in this particular analysis, we can differentiate between 'out-degree' (the number of edges leaving a node; for example, the out-degree of a process is incremented for each artefact it becomes connected to via a *used* edge) and 'in-degree' (the number of edges directed towards a node). In each of the six distributions, the tails can again be well fitted by a power law, with an exponent (DPE) ranging from 1.9 to 4.1. Specifically, from Table 1 it is apparent that the values of DPE in the figure are (a) 2.17, (b) 4.11, (c) 1.86, (d) 3.09, (e) 3.02, and (f) 3.32.

Examining degree distributions by edge type leads to more provenance-specific information, and we highlight some results here. First, examining the number of processes versus the *wasControlledBy* out-degree confirms that each process was controlled by exactly one agent; the plot is not shown since it contained just this one data point (out-degree 1, number of processes 37,931); in Table 1 it is noted that this case has "No power law". Second, examining the number of processes versus the *wasGeneratedBy* out-degree confirms that each artefact was generated by exactly one process. Again the plot need not be shown; here the single data point was out-degree 1, number of artefacts 58,877. This was a gratifying result, as it is always the case in a single account that an artefact is generated by a single process/activity (more generally, different accounts may model what happened from different viewpoints, and the same entity may be recorded as generated from two different processes in two accounts). This confirmation would not be pertinent to normal CollabMap users, but it could be of use to developers

Table 1. Values of the degree-distribution power-law exponent (DPE) for the four types of node inter-connection, when power laws were observed

	used	*wasGeneratedBy*	*wasControlledBy*	*wasDerivedFrom*
in-degree	2.17	3.02	3.32	1.86
out-degree	3.09	(No power law)	(No power law)	4.11

wishing to check the accuracy of the implementation of their software. Third, let us consider the fact that the degree distribution for artefacts is essentially determined by the number of times an artefact is reused. From the distribution in Figure 7(a), we found the average in-degree was 0.80, and the conclusion to draw from this is that each artefact was used slightly less than once, on average. Additionally, the range of in-degrees was 0–35; hence some artefacts were used very heavily while some artefacts were not used at all. The latter are mostly user votes (over 43,000), which were recorded for data verification at a later stage and not currently used in any of the micro-tasks. Artefacts that were used at all were used an average of 3.4 times. Similar analysis applies to the other plots in Figure 7.

4.2 Graph Diameter

Figure 8 plots the evolution of graph diameter (the maximum separation between any two nodes) as more and more processes occur. It shows that graph diameter tended to increase quickly for the first few processes before settling to a stable value. Results over all 5,128 provenance graphs are shown in Figure 9. Growth is rapid until approximately the seventh process; thenceforth there is a slow, approximately linear increase in graph diameter. The plots in their entirety are sub-linear. In comparison, in many graphs the diameter grows approximately logarithmically with the number of nodes [4], which is of course another sub-linear pattern, and hence qualitative similarities exist. We have begun to show that the dynamics of provenance graphs bear some resemblance to those of other networks in the literature.

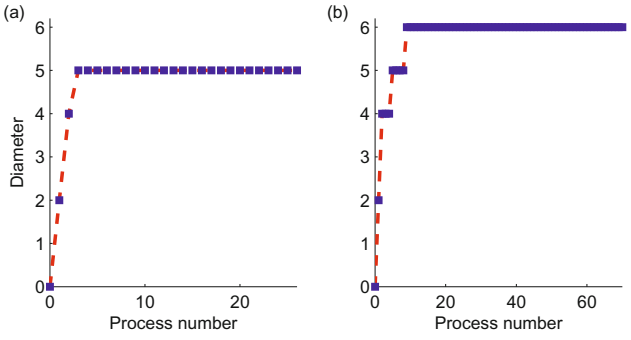

Fig. 8. Plot of diameter versus process number for **(a)** a typical provenance graph, and **(b)** the largest provenance graph

We have also noted a slight difference, one which is due to the type of expansion expected of a provenance graph: the number of artefacts in a chain growing with each process is a complicated function that nonetheless should in many cases contain a small linear term; this in turn leads to the slightly unusual phenomenon of linear growth after a certain number of processes occur. More specifically, revisiting the workflow description in Section 2, consider that the provenance graph depicted in Figure 3 (with a diameter of four) has the capacity to expand downwards through, for example, *wasRevisionOf* edges. If the artefacts downstream are used by processes controlled by agents who have contributed previously to the task, the diameter will not increase, because the agents will have high degree and will act as 'hubs' keeping all nodes within short reach of one another. On the other hand, if new agents control the processes using these downstream artefacts, there is nothing to prevent graph diameter from growing steadily as more and more downstream artefacts appear. Therefore, the linear growth observed in Figure 9 after approximately the seventh process is an indication that, among other things, a fresh supply of agents is readily available, which is the case for crowdsourcing applications in particular.

Recall from Section 3 that the path length between a pair of provenance nodes is measured by the number of directed edges to be traversed in order to travel from one node to another, and the calculation of most path-length data necessitates first ignoring the node pairs with infinite path length between them. Among the remaining node pairs, the maximal finite distance (MFD) between any two processes in a graph was between 1 and 13 edges, inclusive, and the mean was 2.73 edges. The usefulness of this number becomes apparent only when seen in the context of others — namely, the distance required to go from an artefact to a process. In the latter case, the MFD was also 2.73, and hence, the separation statistics were identical in these two cases. This equality is due to the manner in which the CollabMap provenance graphs were created. Under the OPM, two processes are not connected directly but are linked via artefacts. The second process, i.e. the one using the intermediate artefact, will have generated artefacts of its own; hence these artefacts will be separated from the first process by the same distance (i.e. two edges) that exists between the two processes. This is particular to CollabMap, because when an artefact is connected to a process via *wasGeneratedBy*, and that process uses a second artefact, there is always a *wasDerivedFrom* link between the two artefacts. The motif that results (two edges long) is repeated as the provenance graph grows, and hence for any CollabMap graph the pair of MFD values described here are always equal to one another. For example, clearly the values are both 2 for the small graph depicted in Figure 3. In summary, mean separation data can provide a rapid indication of how the provenance graph model was established initially by its designer.

Similarly, the MFD between two artefacts in a given CollabMap provenance graph was found to have a mean of 1.74 edges (range: 1–12), and the same can be said for the distance required to go from a process to an artefact, and for the distance between two artefacts in the corresponding 'data-flow graph' (see Section 3). Again, the separation statistics were identical in these three cases owing to the manner in which the CollabMap provenance graphs were created. The rationale is only a slight variation on the motif described above, and as a specific example of the phenomenon, in the provenance graph depicted in Figure 3 the MFD between two artefacts and the MFD going from

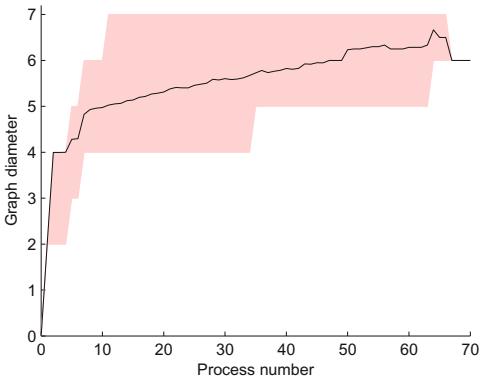

Fig. 9. Evolution of graph diameter with the number of processes (micro-tasks) that have occurred, for up to 5,128 CollabMap provenance graphs. The solid line indicates the mean, and the shaded region indicates the point-wise range of values. The number of graphs available for analysis after each process is given in Figure 5.

a process to an artefact are both equal to 1. In an arbitrary provenance graph outside of the CollabMap project, there may exist a different relationship among the MFDs rather than equality; hence, this relationship provides another measure characterizing the design of a provenance graph model. It is necessary to confirm this by repeating the calculation of MFDs on provenance graphs from other applications.

4.3 Densification

Figures 10(a) and 10(b) are included to show densification — that is, the manner in which the number of edges increases with the number of nodes as a graph grows over time. The two logarithmic plots show only minor deviations from the straight line of a power law, and this pattern was typical among the provenance graphs we examined. The densification power-law exponents for these two selected provenance graphs were 1.33 and 1.23, respectively. Over all 5,128 graphs, the mean exponent was 1.31 with a standard deviation of 0.07, and the range was 1.14–1.59. In comparison, the value seen in other networks is never less than unity in a connected graph [10] and typically falls between 1.0 and 1.7 [9]. The full distribution, given in Figure 10(c), is multi-modal as before; however, we found that restricting the analysis to graphs with a minimum size of 40 led to the emergence of a single peak around 1.3. This peak is close to, for example, the value of 1.26 reported by Leskovec *et al.* [8,9] for a person-to-person recommendation network built from data provided by an online retailer, in which nodes represent users and edges represent recommendations (each time a user purchased a product, they were given the option to send emails recommending the item to friends). More generally, that our results fit in the typical range of 1.0 to 1.7 suggests that provenance graphs grow in a manner that has similarities with other graphs. In addition, the observed standard deviation (0.07) was relatively small, which is related to the fact that the provenance graphs grew in a structured manner with each micro-task.

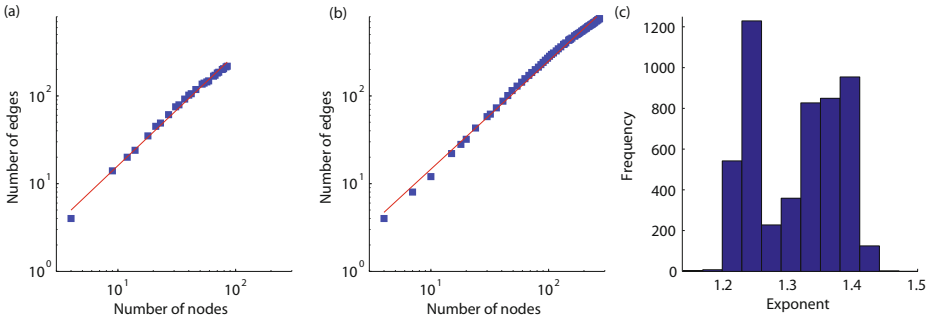

Fig. 10. (a) A plot of the number of edges versus the number of nodes in the provenance graph depicted in Figure 4(a), as it grows. (b) A similar plot for the largest provenance graph. (c) A histogram of the densification exponent a, which is a factor influencing the edge-to-node correlation (ENC), over all 5,128 provenance graphs.

We turn to the values of edge-to-node correlation (ENC), which reflect the densification pattern for particular edge types versus particular node types. Table 2 describes the ENC values among the three node types and the four edge types. In all twelve cases, high values of ENC were observed, which explains the very good line of fit in Figures 10(a) and 10(b). Additionally, there is a deterministic, precisely linear relationship between the number of artefacts and the number of *wasGeneratedBy* edges, or between the number of processes and the number of *wasControlledBy* edges, leading to ENC=1 in either case. This meets with intuition, as each process in CollabMap is linked exactly once to an agent, and (as stated previously) each artefact is generated by exactly one process.

Table 2. Edge-to-node correlation (ENC) coefficients between the number of edges and the number of nodes in a growing graph, averaged over the 5,128 tasks. The three node types are listed on the left and the four edge types are at the top.

	used	***wasGeneratedBy***	***wasControlledBy***	***wasDerivedFrom***
artefact	0.9888	1.0000	0.9929	0.9990
process	0.9948	0.9929	1.0000	0.9894
agent	0.9707	0.9809	0.9807	0.9771

5 Conclusion

In the course of analysing data from a crowdsourcing application, we have highlighted several graph-theoretic metrics to characterize provenance graphs, including DPE, MFD, and ENC. Our first key finding is that CollabMap provenance data possess characteristics similar to those existing in other graphs studied in the literature, including social networks and the World Wide Web [10]. Our second key finding is that our data set is amenable to tools more specific to provenance: our metrics can be used to compare

and classify provenance graphs, to help quickly confirm that provenance was recorded properly, and so on.

The first key finding is important because the similarities we have identified indicate that provenance graphs represent a suitable area of exploitation for network analysis tools concerned with modelling, prediction, and inference which exist already in the literature [7]. For example, since the mid-2000s interest has been growing in 'community detection' — that is, identifying groups of nodes that are more densely linked to each other than to the rest of the network. Users in CollabMap (represented as agent nodes) should not form such communities since tasks are assigned at random, and therefore to the extent that community structure is discerned a pathological case is likely. As an example of such a case, CollabMap users have the option to forgo a task and move on to another, thereby allowing them to focus on particular types of task if desired; hence a group of users could agree among themselves to each skip tasks until they recognized a building or buildings of common interest (for example, in a neighbourhood they disliked). The user group could then corroborate each others' bogus building evacuation routes. Community-detection algorithms such as those based on non-negative matrix factorization [16] could help to alert CollabMap designers to inappopriate levels of community structure within the provenance graph, and thus identify and prevent ill-intentioned collaboration among users. As another example, elsewhere we are in the course of developing a link-inference algorithm based on our results here, to assist with the analysis of incomplete provenance graphs.

The second key finding is important because the set of provenance-specific measures from network analysis so far is useful in its own right, in verification and classification, for example. We have shown how degree distributions can be used to confirm provenance graphs were constructed properly, and the plots in Figure 7 illustrate how further properties in a provenance database can be summarized. Other characteristics we have calculated are the maximum path lengths separating given types of nodes, and densification information. In all of the above, the analysis could have been performed on provenance graphs one at a time rather than on an entire database; a useful application of doing so would be to assist the principled comparison of one provenance graph with another. For example, insofar as our metrics are related to completeness and error probability, they can be used in the process of automated verification of the crowdsourced evacuation routes (e.g. confirming that the editing processes were likely to reduce errors acceptably). In machine-learning terminology, the metrics represent the result of 'feature extraction' and as such they have the potential to help learn the differences between high-error graphs and low-error graphs. In general the metrics are of potential use in future software applications which aim to classify tasks based on their provenance graphs.

Acknowledgements. We gratefully acknowledge funding from the UK Research Council for project 'Orchid', grant EP/I011587/1.

References

1. Altintas, I., Anand, M.K., Crawl, D., Bowers, S., Belloum, A., Missier, P., Ludäscher, B., Goble, C.A., Sloot, P.M.A.: Understanding Collaborative Studies through Interoperable Workflow Provenance. In: McGuinness, D.L., Michaelis, J.R., Moreau, L. (eds.) IPAW 2010. LNCS, vol. 6378, pp. 42–58. Springer, Heidelberg (2010)

2. Batagelj, V., Mrvar, A.: Pajek-program for large network analysis. Connections 21(2), 47–57 (1998)
3. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., Panovich, K., Arbor, A.: Soylent: A Word Processor with a Crowd Inside. In: Artificial Intelligence, pp. 313–322 (2010)
4. Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. Proc. Natl. Acad. Sci. USA 99, 15879–15882 (2002)
5. Clauset, A., Shalizi, C., Newman, M.: Power-law distributions in empirical data. SIAM Review 51, 661–703 (2009)
6. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1(1), 269–271 (1959)
7. Kolaczyk, E.: Statistical Analysis of Network Data. Springer (2009)
8. Leskovec, J., Adamic, L., Huberman, B.: The dynamics of viral marketing. In: ACM Conference on Electronic Commerce (2006)
9. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data 1(1), 2 (2007)
10. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z.: Kronecker Graphs: An Approach to Modeling Networks. Journal of Machine Learning Research 11, 985–1042 (2010)
11. Margo, D., Smogor, R.: Using provenance to extract semantic file attributes. In: Proceedings of the 2nd Conference on Theory and Practice of Provenance, TAPP 2010, p. 7. USENIX Association, Berkeley (2010)
12. Milgram, S.: The small world problem. Psychology Today 1, 61–67 (1967)
13. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., Van den Bussche, J.: The Open Provenance Model core specification (v1.1). Future Generation Computer Systems (July 2010)
14. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 58 (2003)
15. Newman, M.: Networks: an introduction. Oxford University Press (2010)
16. Psorakis, I., Roberts, S., Ebden, M., Sheldon, B.: Overlapping community detection using Bayesian nonnegative matrix factorization. Physical Review E 83(6), 066114 (2011)