Summarising Wireless Network Datasets

Charlotte Knight

September 2019

Contents

1 Introduction

(CURRENTLY JUST DOER DESCRIPTION) With such large quantities of wireless traffic now travelling through networks at ever increasing rates, processing of this data can be challenging. By introducing a summarisation step before any main processing the overall efficiency of information extraction from wireless network datasets may be increased. The aim of this project will be to create a summarised report from large datasets in order to enable more efficient onward processing of the data. This will mean using statistical approaches to maintain some identified information from the dataset while reducing the overall quantity of data that must be stored and processed. The output summaries created may utilise an existing format if my research can identify an appropriate one. My project will produce a command line application involving new approaches to summarisation to run over data collected in the CRAWDAD archives, the approach taken may (in this project or otherwise, dependent on time constraints) also be extended to work on datasets in real-time so as to eliminate the need for storing large datasets before summarisation.

2 CRAWDAD Usage

2.1 Research

When completing this research the focus has been on one particular dataset from the CRAWDAD archive, dartmouth/campus [10]. This dataset was chosen because it is one of the most popular datasets in the archive, having been cited by 374 papers at the time of writing [2]. The most frequently cited dataset however is cambridge/haggle, the reasoning for deciding not to focus on this instead is that the Cambridge dataset is comparatively small in size and would therefore benefit much less from the summarisation which this project hopes to provide.
The papers that have been selected for use in this research were chosen because they all cite the dartmouth/campus dataset. A Google Scholar [1] online search was used to retrieve the most "relevant" papers which used the chosen dataset, from these results the ones which have been most often cited in other work were selected. This selection process found papers which are relevant in the research community. As there are different versions of the dataset the search had to be repeated three times, once using the 2009 dataset, once with the 2007 dataset, and once with the 2005 dataset. For each search the five most cited results have been used. Table 1 shows a summary of the type of information each paper needed to use from the dartmouth/campus dataset. Papers in which the dataset was referenced but ultimately has not been used have been excluded.

2.2 Summary of Results

The usage of the Dartmouth University CRAWDAD dataset is primarily regarding network mobility and social interaction/encounters. As such, the most

		Properties Needed		
Paper	Topic	Device/AP Identification	Time of Transmission	Transmission Quality/Rate
Nextplace: a spatio-temporal prediction framework for pervasive systems, Scellato et al., 2011	Mobility	x	x	
Community-Aware Opportunistic Routing in Mobile Social Networks, Xiao, Wu, and Huang, 2014	Mobility	x	x	
On nodal encounter patterns in wireless LAN traces, Hsu and Helmy, 2010	Mobility	x	x	
Mobility models for systems evaluation, Musolesi and Mascolo, 2009	DTN	x	x	
Large-Scale Synthetic Social Mobile Networks with SWIM, Kosta, Mei, and Stefa, 2014	Mobility	x	x	
WAVEFORM DESIGN AND NETWORK SELECTION IN WIDEBAND SMALL CELL NETWORKS, Yang and Liu, 2014	Mobility	x		x
MAGA: A Mobility-Aware Computation Offloading Decision for Distributed Mobile Cloud Computing, Shi, Chen, and Xu, 2017	Mobility	x	x	
Flow-Based Management For Energy Efficient Campus Networks, Amokrane et al., 2015	SDN	x		x
Human behavior and challenges of anonymizing WLAN traces, Kumar and Helmy, 2009	Anonymizing WLAN Traces	x	x	
Automatic profiling of network event sequences: algorithm and applications, Meng et al., 2008	Profiling of Network Event Sequences	x	x	
Confidentiality of event data in policy-based monitoring, Montanari and Campbell, 2012	Policy-Based Monitoring	x		
Distribution of inter-contact time: An analysis-based on social relationships, Wei et al., 2013	Distribution of Inter-Contact Time	x	x	
Coverage and Rate Analysis for Facilitating Machine-to-Machine Communication in LTE-A Networks Using Device-to-Device Communication, Swain, Thakur, and Chebiyyam, 2017	Machine-to-Machine Communication	x	x	
Balancing reliability and utilization in dynamic spectrum access, Cao and Zheng, 2012	Dynamic Spectrum Access	x	x	
An Online Algorithm for Task Offloading in Heterogeneous Mobile Clouds, Zhou et al., 2018	Offloading	x	x	
State-of-the-Art Routing Protocols for Delay Tolerant Networks, Feng and Chin, 2012	DTN	x		x

Table 1: Table of the properties of CRAWDAD dartmouth/campus data used in various research projects in which it was cited. Papers are ordered by the number of other papers they have been cited by, with the most cited at the top.

often needed information seems to be identifiers for both mobile devices and access points, and the times of connections. I found that the majority of the papers I looked at used the movement [12] or syslog [11] tracesets as these are most tailored towards mobility research.

There are also some less frequent topics of research such as software defined networking and delay tolerant networking using the dartmouth/campus dataset. These uses seem to require a wider variety of information from the data, however these instances are much less frequent than those mentioned above. these less common cases are the only ones which mention bandwidth and quality of connection.

3 Formats

3.1 Existing Formats for Aggregation of Network Traces in Mobility Research

A lack of published information on the intermediate formats used while analysing network traces for mobility research has been found during this research. This is likely due to the encounter data not being the final outcome of the research taking place and therefore not being considered important enough to write up.

Through varied searches of DTN, Mobility, and SDN research I have found only two examples of well documented formats for storing data on device encounters. The first of these, The ONE Simulator [8], uses several reporting options to store device encounter data. The other documented format (from [23]) that was found was an association matrix. These two sources and and analysis of the information found in them is set out in the following section of this report.

3.1.1 The ONE Simulator [8]

The ONE is a simulator which generates data intended to mimic a network of mobile nodes. It then reports this data using various reporting modules, three of these modules focus on data regarding encounters between devices.

The first and most simple of these reports contains information about the dispersion of the total number of encounters experienced by the nodes in the network. It consists of two fields, one containing the number of encounters, and the other containing the count of nodes that have experienced that number of encounters. This contains no information relating to the unique nodes between which the encounters occur, or any temporal information such as duration of the encounters.

The second format provides information on the uniqueness of the encounters recorded, but loses detail about the total number of encounters in the dataset. This format also contains two data fields, one containing the values from 0 to 1000; representing promilles. The second field contains the number of unique pairs encountering with frequency withing the corresponding promille. This has

	Level of Detail (Complete, Most, Some, or None)		
Format	Endpoints	Duration	Frequency
TotalEncountersReport - The ONE [8]	Some	None	Complete
UniqueEncountersReport - The ONE [8]	Some	None	Most
EncountersVsUniqueEncounters - The ONE [8]	Most	None	Complete
Association Matrix [23]	Complete	Some	Some

Table 2: Table of existing formats for storing data about device encounters and the level of detail they contain regarding the unique endpoints and length/frequency of the encounters.

a benefit of being almost static in size as the number of nodes int he system increases.

The final report format from the ONE simulation which has been looked at is a combination of the two previously discussed reports. It has three fields; the first contains an identifier for each node, the second contains the total number of encounters that the node has had, and the third contains the number of unique nodes with which it has had an encounter. This still does not uniquely identify both devices in an encounter, nor does it provide and detail about the duration of the encounters.

3.1.2 Association Matrices

Thakur et al. use an association matrix to record the percentage of time each node spends in an encounter with each other node. A matrix is created for each node, each column in the matrix corresponds to the other endpoint of the encounter, and each row corresponds to a time interval. The entry in each cell represents the percentage of the time interval spent in an encounter with the columns node. This format contains the most information out of all those discussed here, however also takes more space. The space taken will increase at with the square of the number of nodes.

3.2 Summary of Findings

A very brief summary of the detail contained within each of the formats discussed above is given in Table 2. Despite association matrices storing the most useful information, the polynomial increase in size with the number of mobile nodes makes using them potentially ineffective in the context of this project. The aim is to summarize a large amount of data into a smaller, easier to process format. In many cases association matrices would decrease the size of the data, but by a dramatically lesser amount than the other formats discussed here. Additionally it would be possible for the association matrix format to increase the quantity of data; for instance if N nodes had $< N$ encounters each

then the association matrix for each node would include at least one redundant column. The final reporting format discussed under the ONE simulation - EncountersVsUniqueEncounters - avoids this polynomial growth, with its size increasing only linearly with the number of mobile nodes in the network. It provides less complete information regarding the unique par of nodes between which the encounter occurred, there would however be no way to preserve this information while avoiding at least N^2 growth in size.

It seems that the most complete format of those discussed in which to store encounter data while also guaranteeing a reduction in the quantity of data stored would be the EncountersVsUniqueEncounters report format. This format could also be easily modified to add additional fields such as statistics regarding encounter duration. Any additional fields would need to be carefully considered and justified in order to keep the data quantity reduction as high as possible.

4 Design of Code

5 Implementation

6 Outcome/Evaluation

References

[1] URL: https://scholar.google.com/schhp?hl=en&as_sdt=2005&sciodt=0,5.

[2] *About CRAWDAD*. 2014. URL: http://crawdad.org/about.html.

[3] Ahmed Amokrane et al. "Flow-Based Management For Energy Efficient Campus Networks". In: *IEEE Transactions on Network and Service Management* 12 (Dec. 2015), pp. 1–1. DOI: 10.1109/TNSM.2015.2501398.

[4] G. Baudic, Tanguy Pérennou, and Emmanuel Lochin. "Following the Right Path: Using Traces for the Study of DTNs". In: *Computer Communications* 88 (May 2016). DOI: 10.1016/j.comcom.2016.05.006.

[5] Lili Cao and Haitao Zheng. "Balancing reliability and utilization in dynamic spectrum access". In: *IEEE/ACM Transactions on Networking (TON)* 20.3 (2012), pp. 651–661.

[6] Zhenxin Feng and Kwan-Wu Chin. *State-of-the-Art Routing Protocols for Delay Tolerant Networks*. Oct. 2012. URL: http://arxiv.org/pdf/1210.0965.p.

[7] Wei-jen Hsu and Ahmed Helmy. "On nodal encounter patterns in wireless LAN traces". In: *IEEE Transactions on Mobile Computing* 9.11 (2010), pp. 1563–1577.

[8] Ari Keränen, Jörg Ott, and Teemu Kärkkäinen. "The ONE Simulator for DTN Protocol Evaluation". In: *SIMUTools '09: Proceedings of the 2nd International Conference on Simulation Tools and Techniques*. Rome, Italy: ICST, 2009. ISBN: 978-963-9799-45-5.

[9] S. Kosta, A. Mei, and J. Stefa. "Large-Scale Synthetic Social Mobile Networks with SWIM". In: *IEEE Transactions on Mobile Computing* 13.1 (Jan. 2014), pp. 116–129. DOI: 10.1109/TMC.2012.229.

[10] David Kotz et al. *CRAWDAD dataset dartmouth/campus (v. 2009-09-09)*. Downloaded from https://crawdad.org/dartmouth/campus/20090909. Sept. 2009. DOI: 10.15783/C7F59T.

[11] David Kotz et al. *CRAWDAD dataset dartmouth/campus (v. 2009-09-09)*. Downloaded from https://crawdad.org/dartmouth/campus/20090909/syslog. traceset: syslog. Sept. 2009. DOI: 10.15783/C7F59T.

[12] David Kotz et al. *CRAWDAD dataset dartmouth/campus (v. 2009-09-09)*. Downloaded from https://crawdad.org/dartmouth/campus/20090909/movement. traceset: movement. Sept. 2009. DOI: 10.15783/C7F59T.

[13] Udayan Kumar and Ahmed Helmy. "Human behavior and challenges of anonymizing WLAN traces". In: *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*. IEEE. 2009, pp. 1–6.

[14] Xiaoqiao Meng et al. "Automatic profiling of network event sequences: algorithm and applications". In: *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*. IEEE. 2008, pp. 266–270.

[15] Mirko Montanari and Roy H Campbell. "Confidentiality of event data in policy-based monitoring". In: *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012)*. IEEE. 2012, pp. 1–12.

[16] Mirco Musolesi and Cecilia Mascolo. "Car: context-aware adaptive routing for delay-tolerant mobile networks". In: *IEEE Transactions on Mobile Computing* 8.2 (2008), pp. 246–260.

[17] Mirco Musolesi and Cecilia Mascolo. "Mobility models for systems evaluation". In: *Middleware for network eccentric and mobile applications*. Springer, 2009, pp. 43–62.

[18] Anthony J Nicholson and Brian D Noble. "Breadcrumbs: forecasting mobile connectivity". In: *Proceedings of the 14th ACM international conference on Mobile computing and networking*. ACM. 2008, pp. 46–57.

[19] Salvatore Scellato et al. "Nextplace: a spatio-temporal prediction framework for pervasive systems". In: *International Conference on Pervasive Computing*. Springer. 2011, pp. 152–169.

[20] Yan Shi, Shanzhi Chen, and Xiang Xu. "MAGA: A Mobility-Aware Computation Offloading Decision for Distributed Mobile Cloud Computing". In: *IEEE Internet of Things Journal* PP (Nov. 2017), pp. 1–1. DOI: 10.1109/JIOT.2017.2776252.

[21] M. Sun et al. "Efficient Articulation Point Collaborative Exploration for Reliable Communications in Wireless Sensor Networks". In: *IEEE Sensors Journal* 16.23 (Dec. 2016), pp. 8578–8588. DOI: 10.1109/JSEN.2016.2611594.

[22] S. N. Swain, R. Thakur, and S. R. M. Chebiyyam. "Coverage and Rate Analysis for Facilitating Machine-to-Machine Communication in LTE-A Networks Using Device-to-Device Communication". In: *IEEE Transactions on Mobile Computing* 16.11 (Nov. 2017), pp. 3014–3027. DOI: 10.1109/TMC.2017.2684162.

[23] Gautam S Thakur et al. "Gauging human mobility characteristics and its impact on mobile routing performance". In: *International Journal of Sensor Networks* 11.3 (2012), pp. 179–191.

[24] K. Wei et al. "Distribution of inter-contact time: An analysis-based on social relationships". In: *Journal of Communications and Networks* 15.5 (Oct. 2013), pp. 504–513. DOI: 10.1109/JCN.2013.000090.

[25] M. Xiao, J. Wu, and L. Huang. "Community-Aware Opportunistic Routing in Mobile Social Networks". In: *IEEE Transactions on Computers* 63.7 (July 2014), pp. 1682–1695. DOI: 10.1109/TC.2013.55.

[26] Yu-Han Yang and K. J. Ray Liu. *WAVEFORM DESIGN AND NETWORK SELECTION IN WIDEBAND SMALL CELL NETWORKS*. June 2014. URL: http://hdl.handle.net/1903/14834.

[27] Bowen Zhou et al. “An Online Algorithm for Task Offloading in Heterogeneous Mobile Clouds”. In: *ACM Transactions on Internet Technology* 18 (Jan. 2018), pp. 1–25. DOI: 10.1145/3122981.