## HUTCHINSON — Social

## <0 images>

## Social Media Data: Twitter Scraping on NeCTAR

## Hutchinson, J., Hammond, J., Martin, F. and Yazbek, D.

The rapid growth in social media communications research in the last five years has seen the assembly of complexly connected datasets of a scale and scope previously rare in humanities scholarship, highlighting the sociocultural intricacies of follower networks (see Weller et al., 2013). Big data inquiry of social media has also been driven by the emergence of publicly accessible, integrated aggregation, indexing, query, and visualisation approaches (Hansen et al., 2011; Burnap et al., 2013). Indeed, the big data moment has challenged humanities researchers to develop innovative methodologies to extrapolate new social knowledge, and brought humanists and computer scientists together in the pursuit of robust eResearch infrastructure and workflows to support these objectives.

This paper explores a novel social media network analysis (SMNA) research methodology, which resulted in the development of a Twitter visualisation tool for the NeCTAR research cloud. It explores the workflow issues of using large-scale eResearch infrastructure for digital humanities research, and discusses the results of the research program on social media network mapping. In doing so, the authors demonstrate the complexities of interdisciplinary humanities and computer science research collaboration, while revealing new insights made possible through eResearch partnerships.

SMNA as a methodology requires the development of an integrated workflow using National eResearch Collaboration, Tools and Resources (NeCTAR) as the computing service to host both a Twitter scraper and a network visualisation tool. We used the native Twitter application programming interface (API) to maximise flexibility, minimise costs, and reduce reliance on commercial third-party data providers. The workflow included scripting an automated ‘clean-up’ phase so that hundreds of thousands of raw tweets could be gathered and indexed in a useful format. Analysing the characteristics of social networks across large datasets is computationally taxing, and to solve this problem we set up an interactive visualisation program, Gephi, in the NeCTAR cloud environment.

During the discovery and implementation phases of the project we made the following key observations. In Australia, there has been significant investment in eResearch infrastructure, including large-scale storage, data discovery, and high-performance and cloud computer systems. However, these services often have imposing barriers to entry for humanities researchers, who either lack the technological skills, desire, or established collaborative networks to apply these methodologies to their field of research (Meyer and Dutton, 2009). Moreover, eResearch assemblages are generally constructed by computer or physical scientists, who may be unfamiliar with the philosophies and theoretical perspectives of social scientists, and so it is not always clear that these technologies can immediately address the humanities’ ‘big data’ problems (Meade et al., 2013). However, in constructing the NeCTAR-based Twitter research infrastructure and pursuing an extended collaborative approach, the yield from the original research corpus provided more unique and novel results than either discipline could deliver alone. Traditionally, computer science researchers have sought interdisciplinary collaborations in the natural sciences and engineering to find appropriate research problems that balanced scale, complexity, and solvability. The materialisation of big data as a research concern within the humanities now gives software engineers new real-world applications of their techniques, a process that is key to the computer science discipline’s evolution (Hopcroft et al., 2011).

The research presented in this paper demonstrates that early methodological discussions between humanists and computational specialists strengthened the research design, producing collaboratively designed research questions, while avoiding the high knowledge barriers to lay eResearch entry (Goggin et al., 2014). For example, in this research context, which drew on the expertise of University of Sydney media researchers and NSW Intersect’s computer scientists, a series of new research questions emerged while interrogating the initial social mapping results. The development of a novel computational research workflow emerged from the humanities scholars’ interest in understanding the languages, norms, or rules of communication activities within the Twitter social media platform. Following that research focus, the computer scientists explored a deeper understanding of the relationships between individual users, and the positive and negative communicative sentiment expressed in users’ interactions. The collaboration required significant mutual investment in exploring each research contributor’s agencies and abilities, and ongoing calibration of the research design.

The resulting workflow and research tool have enabled a rigorous interrogation of the communication activities of complex follower networks within the Twitter platform, and the evolution of a methodology that addresses ethical concerns about big data research raised by boyd and Crawford (2012) along with Ess (2014). It will also underpin the development of an automated research tool that can be rolled out across a number of humanities research projects, and within virtual lab environments.

**References**

**boyd, d. and Crawford, K.** (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication, & Society,* **15**(5): 662–79.

**Bradley, J.** (2009). What the Developer Saw: An Outsider’s Ciew of Annotation, Interpretation and Scholarship. *New Paths For Computing Humanists,* **1**(1).

**Burnap, P., Rana, O. and Avis, N.** (2013). Making Sense of Self-Reported Socially Significant Data Using Computational Methods. *International Journal of Social Research Methodology, Computational Social Science: Research Strategies, Design and Methods,* **16**(2).

**Ess, C.** (2014). At the Intersections between Internet Studies and Philosophy: ‘Who Am I Online?’ *Philosophy & Technology,* **25**(3): 275–84.

**Goggin, G., Dwyer, T., Martin, F. and Hutchinson, J.** (2014). Finding Mobile Internet Policy Actors in Big Data: Methodological Concerns in Social Network Analysis. Paper presented at *Australasian Association of the Digital Humanities, Expanding Horizons,* Perth, Australia, 18–21 March 2014.

**Hansen, D. L., Shneiderman, B. and Smith, M. A.** (2011). *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann.

**Hopcroft, J. E., Soundarajan, S. and Wang, L**. (2011). The Future of Computer Science. *International Journal of Software Informatics,* **5(**4): 549–65, http://www.ijsi.org/1673-7288/5/i110.htm.

**Meade, B., Manos, S., Sinnott, R., Fluke, C., van der Knijff, D. and Tseng, A.** (2013). Research Cloud Data Communities. *THETA: The Higher Education Technology Agenda 2013,* Hobart, Tasmania, 7–10 April 2013, http://eprints.utas.edu.au/16326/1/THETA\_2013\_Meade\_16326.pdf.

**Meyer, E. T. and Dutton, W. H.** (2009). Top‐Down e‐Infrastructure Meets Bottom‐Up Research Innovation: The Social Shaping of e‐Research. *Prometheus,* **27**(3).

**Weller, K., Bruns, A., Burgess, J., Mahrt, M. and Puschmann, C.** (2013). *Twitter and Society.* Peter Lang Publishing, New York.