Knowledge Management for Development Journal
Vol. 8, Nos. 2–3, September–December 2012, 202–212

Routledge
Taylor & Francis Group

The early results of a social network analysis of the KM4Dev Main Discussion Group

Graham Durant-Law*

HyperEdge Pty Ltd, Post Office Box 3076, Manuka, 3076, Australia

This paper presents the early results of a social network analysis of the KM4Dev Main Discussion Group. Ten complete years of data, and two years of incomplete data, were provided for analysis. Data was in an XML format and required a considerable iterative data cleansing exercise. Ultimately this process left 703 identified individuals in the network. These people comprise the node-set for the public bounded or contained network, for which activity and various network measures can be applied. Gloor's (2006) Contribution Index was used to attribute and partition the network. 113 key participants were identified as being crucial to the health of the active public network; however, this group appears to be in decline. Overall the Main Discussion Group of the KM4Dev community appears to be a 'knowledge seeking' network rather than a 'knowledge sharing' network.

Background

In February 2012 KM4Dev published a worldwide call for proposals to conduct a social network analysis of the community. Social network analysis is a 'know-who' knowledge mapping strategy that reveals the inherent complexity of an organization, and in particular the invisible organization (Farmer 2008). Shadbolt and Milton (1999) suggest that the underlying basis for effective knowledge mapping strategies is based on recognizing that there are different:

- types of knowledge;
- types of expert and expertise;
- ways of representing knowledge; and
- ways of using knowledge.

Social network analysis uses graph theory to detect patterns of social ties and relationships among actors (de Nooy *et al*., 2005) to provide both a visual and a mathematical analysis (Wassermann and Faust 1999, Carrington *et al*. 2005, Scott 2005). In a knowledge management context, the usual intent is to identify those individuals who are the 'central connecters', 'knowledge brokers', and 'boundary spanners' of the organization.[1] The method includes a well-defined set of measures and analysis tools that are used to both describe and understand relational data (Durland and Fredericks 2006).

*Email: graham@hyperedge.com.au

ISSN 1947-4199 print/ISSN 1871-6342 online

http://dx.doi.org/10.1080/19474199.2012.718143
http://www.tandfonline.com

In March 2012 KM4Dev commissioned HyperEdge Pty Ltd, specifically Associate Professor Graham Durant-Law CSC PhD, to conduct a social network analysis of the KM4Dev community. The purpose of the study was threefold, as follows:

- first, to better understand KM4Dev in terms of issues such as identity, relationships, function and role;
- second, to better understand the relationship of SIWA and SAGE to each other and the main discussion community; and
- third, to better understand where KM4Dev sits in comparison with other networks to help KM4Dev appreciate where they are unique, where they are in terms of a traditional life-cycle, and how they might evolve.

The project has three phases, which have some overlaps. In Phase One the Main Discussion Group, and the SAGE and SIWA special interest communities, were analysed. Work on the project began in the last week of March 2012, with the results for the Main Discussion Group being provided to the KM4Dev Core Group on 17 April 2012. Results for the SAGE and SIWA special interest communities were respectively provided on the 14 May and 24 June 2012. Phase One is therefore complete; however, this paper only presents data and analysis for the Main Discussion Group. A separate paper will be presented in due course for the SAGE and SIWA special interest communities.

Phase Two is currently on hold. It involves an online survey and social network of the 'central connecters', 'knowledge brokers', and 'boundary spanners' identified in Phase One, together with some 'lurkers' and low activity members. Depending on the numbers and the type of questions to be asked, the survey tool to be used will be either HyperEdge's own proprietary tool, or Optimice's ONA Surveys.

Phase Three is the reporting phase. As mentioned above, some early findings in Microsoft PowerPoint format have already been provided to the KM4Dev Core Group, along with cleaned data in a Microsoft Excel format.

Raw data

Ten complete years of data, and two years of incomplete data, were provided for analysis. Data was in an XML format and required a considerable iterative data cleansing exercise. The first iteration reduced the dataset to 10,576 rows and seven columns which had to be further cleaned and manipulated. The tool of choice, at least for the initial cleaning and manipulation stage, was Microsoft Excel. Excel has some very good capabilities including a '*=CLEAN'* command to remove non-printable hidden characters that cause problems in analysis tools.

The dataset contained 10,354 posts. 7238 were reply posts. Of these 'Anonymous' posted 1999 replies to 1374 posts. This represents about 18% of all posts. However, it was necessary to remove 'Anonymous' from the dataset, because 'Anonymous' is almost certainly not a single person, and to leave them in would distort the results. Similarly, identified pseudonyms, aliases, and duplicate names, along with 'self-replies' and no answers were removed. Ultimately this process left 703 identified individuals in the network. These people comprise the node-set for the public bounded or contained network, for which activity and various network measures can be applied.

Initial analysis

Figure 1 is a spring diagram of the Main Discussion Group using 10 years of data. NodeXL (2011) was used to produce this map. NodeXL is an excellent tool for early analysis and to produce insights that can be examined later: however, quite obviously it is impossible to interpret from a visual inspection alone! 703 people are represented by the ‘dots’ in this network. There are 2981 unique links and 1770 are reciprocated – that is, there is an arrow at both ends of the link. The maximum geodesic distance is 8. This means that theoretically everyone in the network can reach each other in a maximum of eight steps. Most can reach everyone in three steps: that is, there are three degrees of separation in the network. The graph density is 0.007325. Graph density measures the number of reported ties in the network, compares it to the number of possible ties, and expresses the result as a percentage or ratio. Graph density appears low given this is a ‘knowledge exchange’ network and less than 1% of all possible links are present.

The network is also characterized by an obvious core group occupying the centre of the graph, with individuals with fewer links around the periphery. In order to make sense of the network it is necessary to attribute the dataset. Several options were available, with the most obvious being year and month of posting. These however are link attributes and in the first instance we are seeking some node attributes. The dataset did not come with node attributes, and web-scraping the Ning Group and matching email addresses with those in the dataset was not realistic given there were 2643 people at the time of analysis. Furthermore, even if it was realistic the results would have been incomplete given people had multiple email addresses and had come and gone from the KM4Dev community. To resolve this problem it was obvious that a calculated or derived attribute would have to be applied.

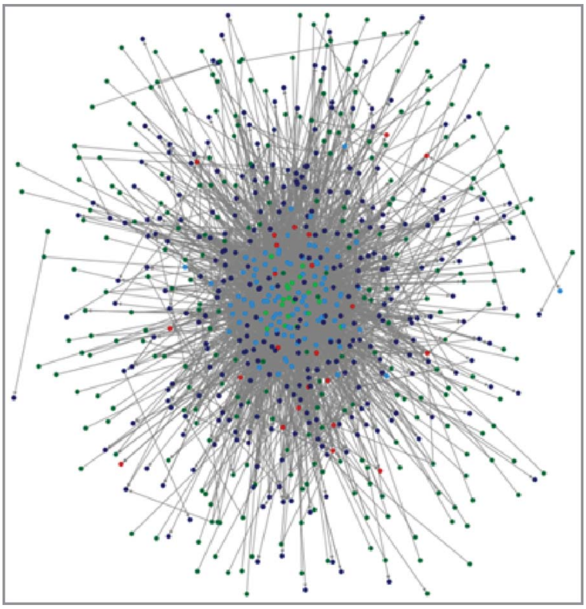


Figure 1. KM4Dev Main Discussion Group 2000–2012.

Gloor's Contribution Index

In order to remove the noise from the network Gloor's (2006) Contribution Index (*messages sent – messages received*)/(*messages sent + messages received*) was applied. Gloor's Contribution Index is interpreted as follows:

- If an individual only sends messages and receives none then their contribution index is +1.000.
- If an individual only receives messages and sends none then their contribution index is –1.000.
- If the communication behaviour is balanced then the contribution index is 0.000.

Coupling the index with the frequency of posting allows an individual's 'role type attribute' to be determined as shown in Figure 2. There are other indices that could be used, including those developed by Hansen *et al.* (2011), but Gloor's Contribution Index is sufficient for this stage of the analysis.

Applying Gloor's Contribution Index identifies 113 key participants, and deeper analysis shows they are active over almost all the years in the dataset. In this analysis high posting frequency is determined by calculating two standard deviations from the mean, and medium posting frequency is calculated as one standard deviation from the mean. The roles are then defined as follows:

- 'No Role' people as low-frequency senders and receivers with a contribution index between –0.499 and +0.499;
- 'Experts' as low-frequency receivers with a contribution index between –0.500 and –1.000;
- 'Envois' as low-frequency senders with a contribution index between +0.500 and +1.000;
- 'Escorts' as medium-frequency senders and receivers, with a contribution index between –0.499 and +0.499; and
- 'Expediters' as high-frequency senders, with a contribution index between 0.000 and 1.000.

Figure 3 is also produced in NodeXL, but this time Gloor's Contribution Index has been applied to data. The power of this diagram is it removes the noise from the network, and we can start to see some network behaviour. The links inside the 'circles' are posts between like roles. Note there are no posts between Experts and only a few between Envois – this

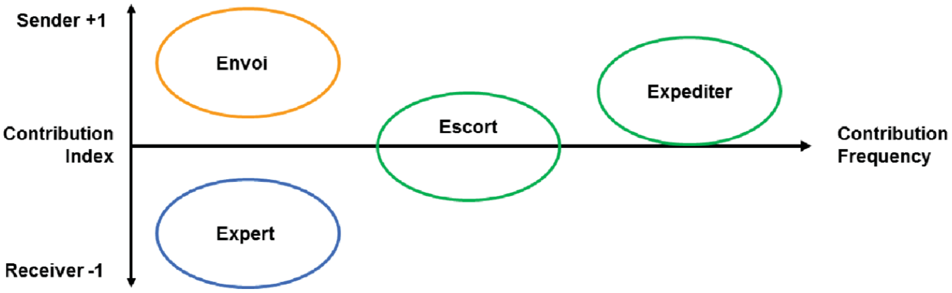


Figure 2. Gloor's Contribution Index.

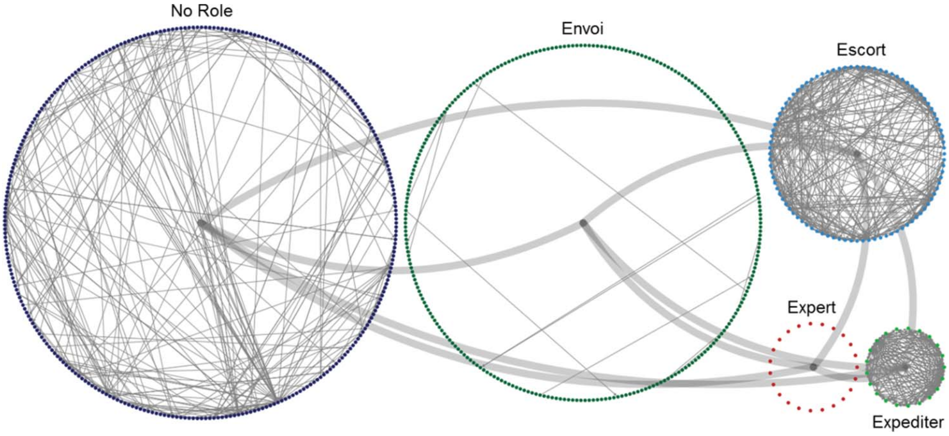


Figure 3. Gloor's Contribution Index applied.

is not unexpected. The thicker curves linking groups are consolidated exchanges between groups. They do not show frequency, or links from one individual to another. Also note the relative density in the Escort and Expediter groups.

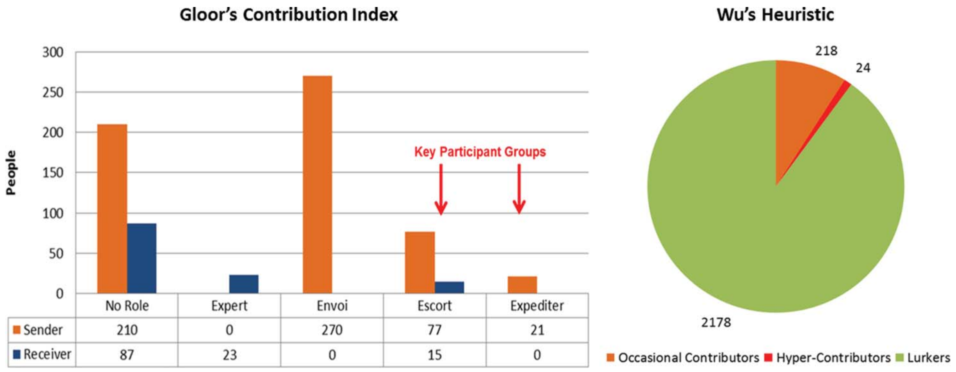
Wu's heuristic

Gloor's Contribution Index provides a way to attribute the network, and a way to examine groups of interest, either as a single group or a combination of groups: however, we also need to know the size of the network before we can make an informed judgment on activity. At the time of data collection in March 2012 there were 2643 people registered on the Ning Group, but this group is different to the dataset, and quite obviously over 10 years the group represented in the dataset had grown and fluctuated in size. It was therefore necessary to estimate the size of the network.

A common heuristic that can be used to determine the size of the network and predict the number of 'lurkers' is the 90-9-1 rule. A study by Wu (2010), using 10 years of data from more than 200 online communities, found that:

- 90% of all users are 'lurkers' who don't actively contribute.
- 9% of all users are 'occasional contributors' providing less than 50% of the content.
- 1% of all users are 'hyper-contributors' providing greater than 50% of the content.

Figure 4 presents data for the Main Discussion Group. Using Wu's heuristic the predicted size of the Main Discussion Group is 2420 people. This is very close to the 2643 people registered on the Ning Group at 31 March 2012, but the difference is statistically significant. Note there is a very close correlation between Gloor's Expediters and Wu's Hyper-Contributors. This difference is not statistically significant. The difference appears to be in the Escort Group, which equals 92 people, compared to Wu's predicted value for Occasional Contributors of 218. The difference is difficult to interpret with confidence, especially given Wu's approach identifies three groups and Gloor's five, but it appears there are far fewer medium-frequency contributors in the KM4Dev Main Discussion Group than expected, and that most of these are seeking knowledge. It is also



	No Role	Expert	Envoi	Escort	Expediter
🟧 Sender	210	0	270	77	21
🟦 Receiver	87	23	0	15	0

Figure 4. Glooris Contribution Index and Wu's heuristic compared.

likely that 10 years of data gives some false reads in that some people who contributed in 2005 are almost certainly not contributing in 2011. This is why a temporal analysis is also required.

A temporal analysis

Figure 5 shows the number of active people in the network with 'Anonymous', identified pseudonyms, aliases, and duplicate names removed. Note the growth of the network and the peak activity in 2008. Further analysis indicates that the discussion group is most active in February and October, with most posts occurring on a Wednesday. The differences are statistically significant. Deeper analysis shows most posts occur between 10am and 2pm Greenwich Mean Time.

More interestingly, note there are never more than 300 active people posting to the network in any given year. Dunbar's (2010) and Wellman's (2011) Numbers are of interest.

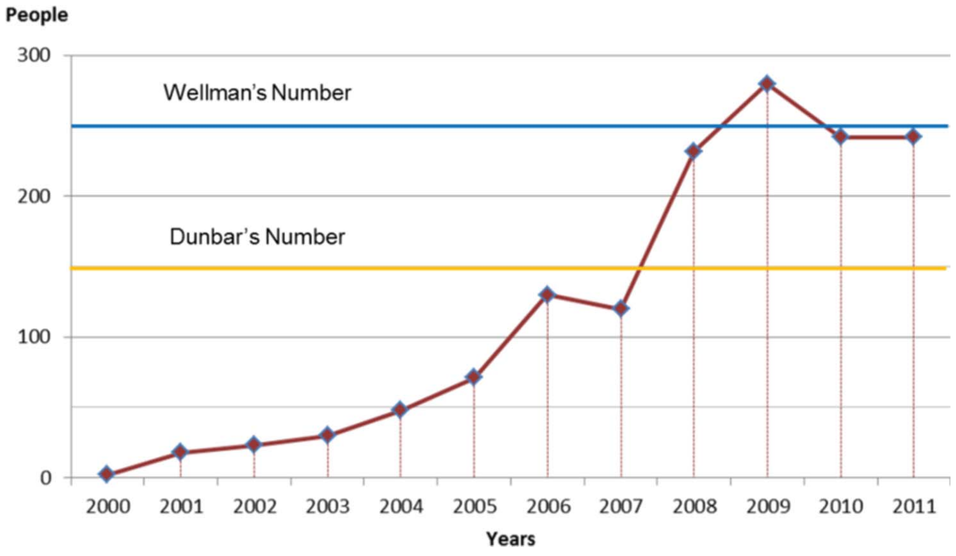


Figure 5. Active people in the Main Discussion Group 2000–2011.

Dunbar's Numbers are an indicator of meaningful strong relationships and the maximum effective number of people in a network. The usually accepted number is 150. There is a mega-band number of around 500, and an upper limit of about 1500. Wellman's number is larger because it takes into account electronic connectivity inherent in the networked era (Wellman 2011, Rainie and Wellman 2012), although the number plotted is a lower order number. Of interest, for the previous four years the active or public network appears to have stabilized around Wellman's number. The literature suggests growth beyond this number may be difficult, unless a number of smaller sub-groups or special interest groups are formed, or the group is actively nurtured.

Partitioning the network

Taking into account the Gloor's Contribution Index analysis and the temporal analysis, it makes sense to focus the analysis on the Escorts and Expediters in the period 2008 to 2011 inclusive. Figure 6 shows the Escort and Expediter networks for 2008 to 2011. In all cases the networks were produced in NetMiner 4 (Cyram 2011), and the Expediters are shown as the hub of the network. However, each individual map does not include the previous year's links: in this sense it is a temporal representation.

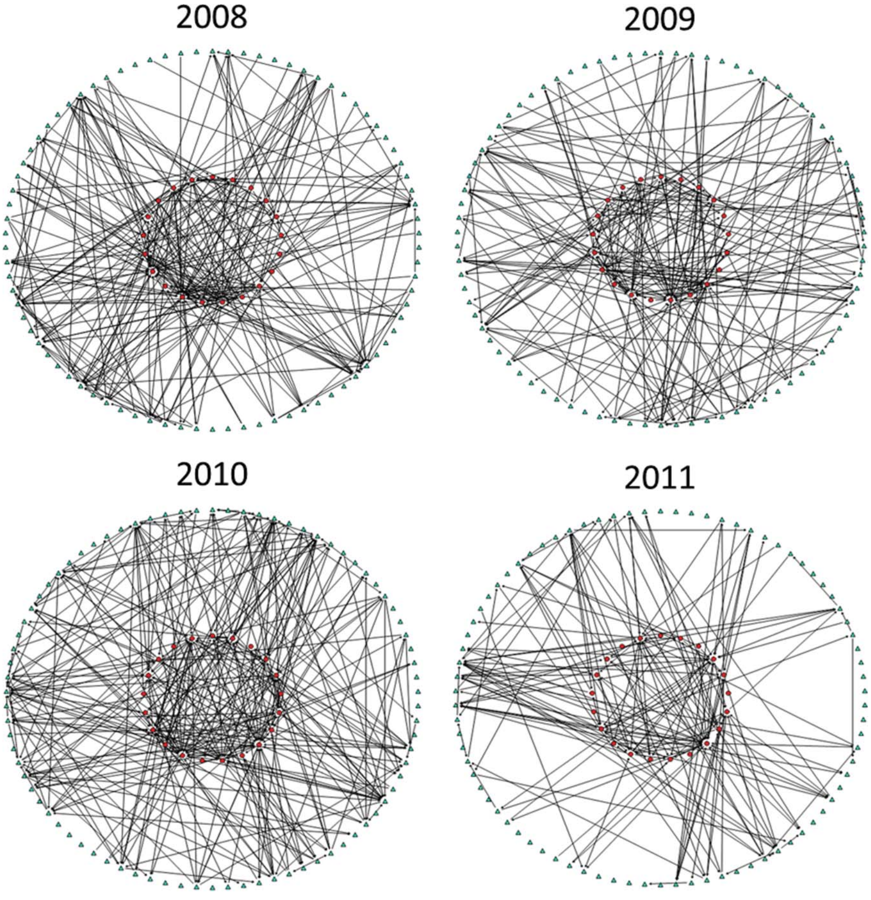


Figure 6. Escort and Expediter networks 2008–2011.

There are 103 people in the complete 2008 to 2011 network. There are 631 unique links and 354 reciprocated links. The maximum geodesic distance is 4. This means that theoretically everyone in the network can reach each other in a maximum of four steps. Most can reach everyone in two steps: that is, there are two degrees of separation in the network. The graph density is 0.093756. Graph density measures the number of reported ties in the network, compares it to the number of possible ties, and expresses the result as a percentage or ratio. Graph density appears low given this is a 'knowledge exchange' network, this is the medium and high-frequency posters, and less than 9% of all possible links are present.

What is striking is the dramatic decline in participation in 2011. Expediters remained in the network and connected; however, 31 Escorts left, or did not participate in, this group. This represents almost 50% of that group. Given Escorts are the 'pivot' group of the KM4Dev Main Discussion Group this may be a weak signal heralding a decline in overall activity in the discussion group. One possible explanation is the decline coincides with the uptake of other social media such as the Ning Group, but deeper analysis does not at this stage support this contention. Further analysis is required.

Discussion

The strength of social network analysis is its ability to show the social, or collective, nature of knowledge within a specific context. It shows the '. . . *actual flow of work and the interlocking roles that often determine the flow of knowledge in organizations*' (Johnson, 2009, e-loc 2609-11). According to Kilduff and Tsai (2005) understanding knowledge networks has an emancipatory effect in the workplace because people become aware of otherwise unknown opportunities and constraints. However, the social network analysis approach has several criticisms. According to Borgatti and Foster (2003) the approach is often criticized for privileging method over theory. A similar criticism is extended to the methodological preoccupation with structure (Monge and Contactor 2003), and that the approach necessarily focuses on the network and its consequences rather than the causal factors (Borgatti *et al*. 2009). It therefore is not dynamic. This analysis suffers from these problems. In particular there is insufficient data to identify causal factors, which means the analysis requires considerable local knowledge to be applied.

A further criticism is the boundary specificity problem. In a research situation just where the system boundary is set drives the collection and sampling techniques to be employed. Where possible it is desirable to collect data for a complete network, but defining just what a complete network is difficult (Wiig 2004), and necessarily confines the outcome. Even in an analysis for management purposes defining the system boundary can be difficult. Johnson (2009, e-loc 1054-56) sums up the problem eloquently in the following passage:

> Perhaps the best-known, and at times most difficult, issue associated with the configuration of networks is where to draw the boundaries around them. This is especially problematic since boundaries imply some discontinuity in relationships; that relationships across boundaries are in some sense qualitatively different than those within the network's boundary.

The problem is if the complete network is not captured there are questions of validity and reliability, and just how representative of the actual networks the knowledge maps are (Marschall 2007). Again this analysis suffers from these problems. The paucity of node attribute data required attributes to be calculated. These derived attributes may not be entirely accurate, and when the network is partitioned may have set artificial boundaries.

On the other hand, the analysis has revealed a number of interesting insights. For example, it is quite clear that comparative to other online communities the KM4Dev Main Discussion Group is relatively quiet and is characterized by low reciprocity. Indeed 69% of all posts are not answered! Even allowing for the broadcast nature of some posts, this figure appears unusually high, and is characteristic of a knowledge-seeking network, rather than a knowledge-sharing network. This would seem to be at odds with the charter of KM4Dev. Deeper analysis reveals this not the case for some of the special interest communities. SAGE in particular has the exact opposite pattern of reciprocity, with 76% of all posts receiving a reply. Perhaps this reflects the homogeneity and small size of the SAGE special interest community?

There are two other important insights, which are closely related. First, posting activity appears to have peaked with the number of active people in the KM4Dev Main Discussion Group stabilizing around Wellman's Number of about 250 people. Second, the active group has a nucleus that has not changed much over the years. Many of these people are, or have been, part of the KM4Dev Core Group for many years. Indeed if some of these people were to stop contributing or participating the KM4Dev Main Discussion Group would likely fracture. This is more so the case for the special interest communities which are centred on one or two key players. These results suggest the future stability of the KM4Dev Main Discussion Group lies in encouraging reciprocal activity and recruiting new participants for leadership roles: both activities are easier said than done! Some actions might include:

- soliciting network weaving and engagement ideas from the Escort, Expediter and No Role groups, given their greater participation;
- consciously closing network triangles within the Envoi and Expert groups to facilitate greater collaboration;
- designating several 'network weavers' from within the Escort and Expediter groups whose task is to connect with the Expert and Envoi groups, and in turn connect them to the appropriate people; and
- identifying and enacting a small project that will weave the people on the edge of the respective groups together.

Conclusion

This paper has presented the early results of a social network analysis of the KM4Dev Main Discussion Group. After data cleaning there was 703 identified individuals in the network. Of these 113 key participants were identified as being crucial to the health of the active public network; however this group appears to be in decline with 31 people either leaving the group or not participating in 2011. Overall the Main Discussion Group of the KM4Dev community appears to be 'knowledge seeking' network rather than a 'knowledge sharing' network.

There are however some limitations to the study and this paper. First, and most significantly, data has been treated as 'deterministic'. That is, measurements have been applied to the KM4Dev Main Discussion Group as if it were in a 'final' or 'equilibrium' network state. Clearly this not the case as the network is constantly changing and evolving. Local knowledge should therefore be applied to the interpretation of data.

Second, the diagrams presented in this document are drawn from filtered data. Data can be filtered in multiple ways resulting in differing diagrams. It is also possible to present other forms of diagrams from the same filtered data. Indeed collectively in the reports provided to the KM4Dev Core Group there are over 200 visualizations. Most make use of network metrics such as degree, closeness, betweenness, and eigenvector centrality that are not discussed in this paper. These measures allow for the identification of specific people who are key to the network, but are limited by the lack of attribute data to enhance the analysis.

Finally, the analysis is somewhat one-dimensional. All we have looked at is links between posters, with no regard to the content of the post. This means a broadcast message with a reply of 'I will attend' has the same weight as a knowledge sharing question and response. This study would be enhanced considerably by a content analysis of the posts, which would likely provide a number of additional insights. However, a content analysis is not a trivial matter and must be weighed against the time and resources required to conduct the analysis, and triangulate the results with this study.

Acknowledgements

I wish to acknowledge Pete Cranston, Natalie Campbell, Riff Fullan, Ewen Le Borgne, and Jaap Pels for their help in supplying and interpreting data. I also thank Sarah Cummings for the opportunity to publish this paper.

Note

1. A central connecter is 'someone who is highly connected to many others in the network, who may be either a key facilitator or a gatekeeper'; a broker is 'someone who communicates across sub-groups'; and a boundary spanner is a 'person who connects a department with other departments' (Anklam 2005, p. 344).

Notes on contributor

Graham Durant-Law (CSC, PhD) is the owner and chief scientist of HyperEdge Pty Ltd. He is an expert in social and organizational network analysis, and developed the business network analysis^{TM} modelling methodology. He is also an acknowledged thought leader in knowledge management, and has been one of the international adjudicators for the Singapore Knowledge Management Excellence Awards for the past six years. Graham holds numerous academic qualifications and has won prizes for academic achievement, as well as awards for innovation and practice. He is an Adjunct Associate Professor at the University of Queensland and the University of Canberra, and a Senior Lecturer at the Australian National University. He was awarded a prize for stakeholder management in the Project Management Institute's Project Manager of the Year Award in 2010. Graham is passionate about building superior organizations using evidence-based methods, and maintains a blog called *Knowledge Matters* with this theme.

References

Anklam, P., 2005. Social network analysis in the KM toolkit. *In*: M. Rao, ed. *Knowledge management tools and techniques*. Oxford: Elsevier Butterworth Heinemann, 329–346.

Borgatti, S. and Foster, P., 2003. The network paradigm in organizational research: a review and typology. *Journal of Management*, 29 (1), 991–1013.

Borgatti, S, Mehra, A., Brass, D. and Labianca, G., 2009. Network analysis in the social sciences. *Science*, 323 (5916), 892–95.

Carrington, P., Scott, J., and Wassermann, S. eds., 2005. *Models and methods in social network analysis*. Cambridge: Cambridge University Press.

Cyram, 2011. *NetMiner® 4*, Cyram Co. Ltd, Seoul, South Korea.

de Nooy, W., Mrvar, A., and Batagelj, A., 2005. *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Dunbar, R., 2010. *How many friends does one person need? Dunbar's number and other evolutionary quirks*. London: Faber and Faber.

Durland, M. and Fredericks, K., eds., 2006. *Social network analysis in program evaluation*. Minnesota: Wiley.

Farmer, N., 2008. *The invisible organisation. How informal networks can lead organizational change*. Farnham: Gower.

Gloor, P., 2006. *Swarm creativity: competitive advantage through collaborative innovation networks*. Oxford: Oxford University Press.

Hansen, D., Shneiderman, B., and Smith, M., 2011. *Analyzing social media networks with NodeXL. Insights from a connected world*. Burlington: Elsevier.

Johnson, D., 2009. *Managing knowledge networks*. Singapore: Cambridge University Press.

Marschall, N., 2007. *Methodological pitfalls in social network analysis. Why current methods produce questionable results*. Milton Keynes: Lightning Source UK Ltd.

Monge, P. and Contactor, N., 2003. *Theories of communication networks*. New York: Oxford University Press.

Kilduff, M. and Tsai, W., 2005. *Social networks and organisations*. London: Sage.

Rainie, L. and Wellman, B., 2012. *Networked: the new social operating system*. Cambridge, MA: MIT Press.

Scott, J., 2005. *Social network analysis: a handbook*. 2nd ed. London: Sage.

Shadbolt, N. and Milton, N., 1999. From knowledge engineering to knowledge management. *British Journal of Management*, 10, 309–322.

Wassermann, S. and Faust, K., 1999. *Social network analysis*. Cambridge: Cambridge University Press.

Wellman, B., 2011. Is Dunbar's number up? *British Journal of Psychology*, 103, 176–6.

Wiig, K., 2004. *People-focused knowledge management*. Oxford: Butterworth-Heinemann.

Wu, M., 2010, *The 90-9-1 rule in reality* [online]. 18 March, Lithium. Available from: http://lithosphere.lithium.com/t5/Building-Community-the-Platform/The-90-9-1-Rule-in-Reality/ba-p/5463 [Accessed 16 July 2012].