Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2009 Proceedings

Americas Conference on Information Systems (AMCIS)

2009

The Small Worlds of Wikipedia: Implications for Growth, Quality and Sustainability of Collaborative Knowledge Networks

Myshkin Ingawale
Indian Institute of Management, Calcutta, myshkinonline@gmail.com

Amitava Dutta
George Mason University, adutta@gmu.edu

Rahul Roy
Indian Institute of Management, Calcutta, rahul@iimcal.ac.in

Priya Seetharaman
Indian Institute of Management, Calcutta, priya@iimcal.ac.in

The Small Worlds of Wikipedia: Implications for Growth, Quality and Sustainability of Collaborative Knowledge Networks

Myshkin Ingawale
Indian Institute of Management, Calcutta
myshkinonline@gmail.com

Amitava Dutta
George Mason University
adutta@gmu.edu

Rahul Roy
Indian Institute of Management, Calcutta
rahul@iimcal.ac.in

Priya Seetharaman
Indian Institute of Management, Calcutta
priya@iimcal.ac.in

ABSTRACT

This work is a longitudinal network analysis of the interaction networks of Wikipedia, a free, user-led collaboratively-generated online encyclopedia. Making a case for representing Wikipedia as a knowledge network, and using the lens of contemporary graph theory, we attempt to unravel its knowledge creation process and growth dynamics over time. Typical small-world characteristics of short path-length and high clustering have important theoretical implications for knowledge networks. We show Wikipedia's small-world nature to be increasing over time, while also uncovering power laws and assortative mixing. Investigating the process by which an apparently un-coordinated, diversely motivated swarm of assorted contributors, create and maintain remarkably high quality content, we find an association between Quality and Structural Holes. We find that a few key high degree, cluster spanning nodes - 'hubs' - hold the growing network together, and discuss implications for the networks' growth and emergent quality.

Keywords

Network Theory, Collaborative Knowledge Networks, Small-Worlds, Wikipedia

INTRODUCTION

In recent times, growth dynamics of collaborative knowledge networks like Wikipedia is attracting attention of IS researchers (Parameswaran and Whinston, 2007; Dutta et. al, 2008). Wikipedia, a free user-led online encyclopedia, does not require users to even register to edit articles. This openness to new users has been cited as both a source of strength and weakness (Hafner, 2006). While "peer production" encourages content generation on a massive scale, at the same time, such openness may increase vulnerabilities to website defacing, destruction of intellectual property, and general chaos (Wagner and Majchrzak, 2007). An interesting research puzzle then is how Wikipedia, growing exponentially in users and content since 2002, has maintained a remarkable level of quality (Chesney, 2006) and generally been highlighted as a success story of low-cost collaborative knowledge systems (Kittur et al, 2007). In this work, using a longitudinal network analysis approach on Wikipedia's knowledge networks, we have attempted to address two specific research issues: a) What are the growth dynamics of Wikipedia? How do its knowledge networks form and grow? and b) What is the association, if any, between the *quality* of content, and the *structure* of the network of interactions in such an open collaborative environment?

While contributing to network theories that have emerged in organizational literature over the past few decades, a better understanding of these two issues would help managers make more informed decisions and develop more effective strategies for design and deployment of collaborative knowledge systems for their own organizations.

The Network Paradigm in Organizational Science

Cast in different styles of argument (e.g. Coleman, 1990; Burt, 1992), the network paradigm in organizational science is a metaphor about *advantage*. It emphasizes the advantages that accrue to nodes[1] as a function of their *position* in the network of interactions, complementary to any *individual–level* attributes that might provide advantages[2]. Schilling and Phelps (2004) provide a generic definition of a knowledge network as "an interconnected set of nodes that receive, store, process and/or transmit information. Nodes may be passive repositories or relays for information, or active agents that search for, integrate, process or broadcast information. The connections between these nodes are referred to as links, and may represent relations (e.g., semantic associations), communication paths (e.g., patterns of regular interaction), or other conduits for information." Two commonly cited mechanisms that yield advantage are *network closure* and *brokerage across structural holes*[3]. The latter is based on the rationale that in a network, individual nodes (like node 1 in the user network of Figure 1c) can benefit from serving as intermediaries between others who are not directly interacting. Through such intermediation, they potentially can broker the flow of information and synthesize ideas arising in different parts of the network. In knowledge networks, nodes spanning such *structural holes* have been found to exhibit positive associations to performance and knowledge diffusion (Burt, 2005). A network with very few structural holes will have a short average path-length[4]. Network closure, on the other hand, implies a dense mesh of local connections, manifested itself by a property called *clustering* i.e. the formation of clusters or *cliques* that are tightly connected internally, but only sparsely connected externally. Many researches have emphasized the role played by this local density in building trust and acting on any available knowledge (Borgatti and Pacey, 2003). Network closure has also been found to correspond to advantages in terms of information-transfer, error-free performance and higher rates of knowledge diffusion (Cowan, 2004).

Small-World Theory

It had been believed for a long time, that short path-lengths and high clustering could not co-exist simultaneously in the same network. This was based on the then dominant random graph models of Erdos and Renyi (1959). However, relatively recent discoveries have revealed that there is a class of networks – called 'small-worlds' - that display both low average pathlengh and high clustering. The first and most popular manifestation of "small worlds" is the "six degrees of separation" concept, uncovered by the social psychologist Stanley Milgram (1967), who concluded that there was a path of acquaintances with typical length about six between most pairs of people in the United States. The seminal work by Watts and Strogatz (1998) found that the small world property appears to characterize many other real-world complex networks. Recent work (Cowan, 2004; Schilling and Phelps, 2004) suggests that small world properties of low average path length and high clustering enable knowledge networks to simultaneously achieve great reach and high bandwidth, accelerating the rate of knowledge creation and increasing the likelihood of acting upon this knowledge.

However, these propositions have rarely been verified through longitudinal studies, presumably owing to the lack of high quality and quantity time-series data on interaction. In a traditional organizational setting, where knowledge networks are often informal and/or transient and hard to monitor, acquiring this data can be a laborious and time-consuming process, even if a large enough volume in a usable form were to exist in the first place. However, a recent paradigm shift, variously called the emergence of the 'second economy' or 'the shared economy' or the 'peer-to-peer economy' (Benkler, 2007) has shifted collaborative knowledge creation outside the boundaries of the traditional firm setting and made available in the public domain quantitatively and qualitatively rich datasets on collaborative production of knowledge.

The rest of this paper is structured as follows: We describe the methodology and dataset used. Then, we present our findings related to the observed small-world characteristics and network association with quality. To uncover the growth dynamics, we present the temporal evolution of the degree distribution, degree correlation and growth of giant component. The 'Interpretation and Discussion' section presents our analysis and implications. We conclude with limitations of this study and directions for future work.

---

[1] Depending on the scale of analysis employed, nodes may represent individuals, departments, organizations, etc.

[2] This stems from a broader discourse in sociology, centered on the view of Social capital as a contextual complement to Human capital. Refer Burt (2005) for a review.

[3] A Structural Hole is defined in terms of the absence of connectivity between two internally connected parts of the network

[4] Pathlength of a node is the minimum number of edges that have to be traversed to reach it from any randomly picked node in the network. The average pathlength of the network is the average pathlength across all nodes.

METHODOLOGY

We use the methodological approach of social network analysis (Wasserman and Faust, 1994). The rationale for this choice is based on our representation of Wikipedia as a *knowledge network* and grounded in the discourse of knowledge as *emergent* from a 'web of relationships' that are the result of *interactions* (Stacey, 2000). Kakihara and Sorensen (2002) summarize this discourse and point out that the network of interactions *is* the knowledge network. While Wikipedia has previously been represented as a network of articles, connected by *hyperlinks* (e.g. Buriol el al, 2006), modeling of the *knowledge* network in its entirety (as opposed to purely its *information architecture*) requires capturing the *interactions* – conflict, negotiation, collaboration - that take place behind the scenes and create the (*misleadingly serene and static!*) 'final' visible article. In Wikipedia, there is an intricate web of relationships born from interactions – between contributors, between articles and between articles and contributors that *shapes* and is *shaped by* the 'final' article: This recursive process suggests a 'duality of technology' (Orlikowski, 1992), where 'agency' and 'structure' are not independent. Wiki 'reality' i.e. the set of 'final' articles, is continually reconstructed by users within a complex network of interactions. Consider the basic use-case in the Wiki publication model: "User edits article". In doing so, the user establishes an interaction tie with a) The previous contributor whose contribution was modified and b) All the other previous contributors, whose contributions, by virtue of not being modified, received a tacit acceptance and approval. In an interaction graph (let us call this 'the user network'), the user is now linked to all the previous contributors of the article. In the user network, each node denotes a user and each edge between 2 users denotes the presence of at least one common article they have contributed to. Similarly, there exists another graph – the article network – where each node represents an article and each edge between two articles denotes the presence of at least one common contributor.

Dataset

We have used the Cebuano language Wikipedia meta-revision history XML file, downloaded from http://download.wikimedia.org/ (29th September 2008). After filtering out robot and non-content entries, it had 1966 articles and 1331 contributing users. There are two advantages to using Cebuano Wikipedia over a much larger language version, say the English Wikipedia: Firstly, the dataset is sufficiently small for us to visualize in one screenshot: This makes it much easier to understand and internalize the formation process. Secondly, it is sufficiently young for us to be able to put the results in a contemporary context, without having to account for technology and major policy changes.

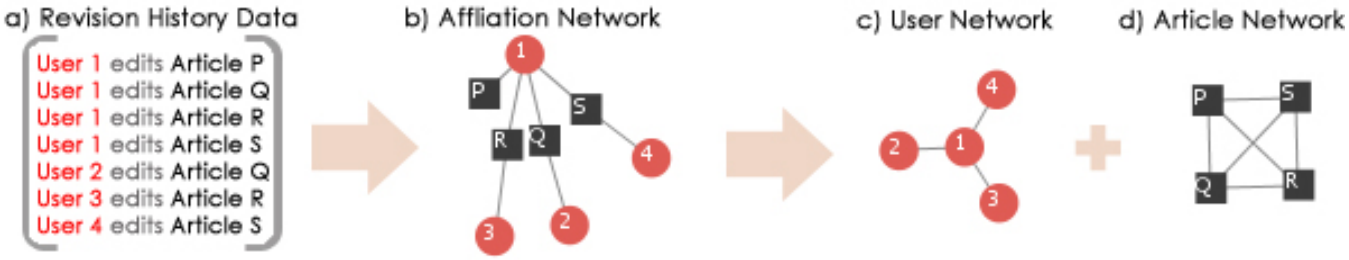

Figure 1. Illustration of how User and Article Networks are generated

Method

We used the NetLogo[5] 4.0.2 development environment to convert the data into graph form. The filtered and parsed meta-revision history data-dump of Cebuano Wikipedia contains information of the type 'user x" edited "article y" at "time t", for every contribution, for every article and as such, constitutes a veritable 'fossil trail' to take us back to relive the dynamics of interaction. First, a multimodal graph – often classed as an *affiliation network* in the literature – was generated from the revision history data. It has two types of nodes: users and articles. This was then converted to the User Network and the Article Network[6], as shown in Figure 1. We generated the networks from the very first contribution i.e. from time zero. Structural properties and visualizations were recorded annually, on 30 Sept of every year i.e. 2005, 2006, 2007 and 2008

[5] NetLogo is a multi-agent programming language and modeling environment for simulating natural and social phenomena. It is particularly well suited for modeling complex systems evolving over time. (Tisue and Wilensky, 2004).

[6] The User and Article network capture complementary but subtly distinct aspects of interaction. The User Network captures *collaboration* information i.e. which users are co-contributors, by virtue of having edited common articles, whereas the Article Network captures *thematic* information i.e. which articles are likely to be based on a common theme or subject area, by virtue of having been edited by common users. Even though obtained from the same source affiliation network, they *may* have markedly different structures, as Figure 1 above illustrates.

Description of Measures

The **Clustering Coefficient** is a measure of how well connected the neighborhood of the node is. If the neighborhood is fully connected, the clustering coefficient is 1 and a value close to 0 means that there are hardly any connections in the neighborhood (Watts, 1999). It is calculated as the ratio of *actual links* to *all possible links* in a node's neighborhood. The clustering coefficient for the whole network is the average of the clustering coefficient for each node. A fully connected neighborhood implies that it forms a cluster that is densely connected internally but only sparsely connected externally. As such this would be indicative of high *network closure.*

The **Pathlength**[7] of a node is "the minimum number of edges that have to be traversed to reach it from any randomly picked node in the network". It follows that nodes of shorter pathlengths are more 'central' in the network, are likely to lie on many 'shortest paths' between nodes and also span more structural holes. The average pathlength of the network is the average of pathlengths of all nodes.

The **Small-World Characteristic Q**, as proposed by Davis et al (2003) and Uzzi and Spiro (2005) is calculated as the ratio

$$Q = \frac{\text{Actual Clustering Coefficient / Clustering Coefficient of Random Network}}{\text{Actual Pathlength / Pathlength of Random Network}}$$

A Q value much larger than 1 indicates that the network is a small-world. The larger the Q-value the more the small-world characteristics of the network. The random networks used as bases for comparison at each flag-point, are generated, as the unipartite versions of a bipartite network with the same number of user nodes, article nodes and links as the corresponding Wikipedia networks[8].

Basic information about *growth* is centered around the evolution over time of the degree (i.e. the *number of edges*) of its nodes. In this regard, degree distribution and degree correlation are very important metrics.

Degree Correlation is the propensity of nodes of similar degree to connect to each other preferentially (Newman, 2002). Common examples of networks with degree correlation are science co-authorship networks (a scientist with many publications would be more likely to collaborate with another high-publication scientist than with any randomly picked scientist). Degree correlation is defined by the following equation

$$k_{nn}(k) = \sum_{k'} k' \Pr(k' \mid k)$$

Pr (k'| k) is the conditional probability that the vertex with degree k is adjacent to the vertex with degree k'. According to this property, networks are said to exhibit *assortative* mixing (or positive correlation) if nodes of a given degree tend to be attached with higher likelihood to nodes with similar degree.

Degree Distribution is the spread in node degree. It is characterized by a distribution function P(k), which gives the probability that a randomly selected node has exactly k edges. While a random network has a uniform spread in node degree, many real-world networks have been observed to have a scale-free degree distribution[9], which, unlike a normal distribution, is characterized by the presence of a small number of nodes with a large number of links and a large number of nodes with a small number of links. It is formed by a 'rich get richer' mechanism whereby as new nodes are added, they attach themselves to old nodes with a probability proportional to the old node's degree (Barabási and Albert, 1999).

As the network grows, the metric '**Size of the Giant Component**' (i.e. the *largest connected chain of nodes*), provides important information about the macro-level network connectivity.

---

[7] Pathlength throughout this text implies *pathlength within the largest connected component.* This is a necessary adjustment because nodes without a path to each other will have pathlength infinity.

[8] We do not directly construct two unipartite base random networks, so as to protect against the problem identified with many real bipartite or affiliation networks - the bias of artificial over-reporting of numerator and under-reporting of denominator in the Q-ratio. We follow the procedure proposed by Newman (2002), of generating a random bipartite network first and making the adjustment of using two rather than one distribution of links (the number of users per article and the number of articles per user) during the construction of the base random network, then converting it into two random unipartite networks.

[9] In such networks, the degree distribution follows a power law relationships, given by $\mathbf{P(k) \sim k^{-\gamma}}$.

RESULTS

As Figure 2, 3 and 4 illustrate, the Article and User Networks grew non-linearly from a very low article and user count, with only a few common edges in 2005, to a much more populated and complex structure, with interlocking strands of clusters in 2008. A dense inner core of highly connected nodes, was seen to form in both the Article and the User Networks, but was much more noticeable by 2008 in the Article Network (Figure 3), which implies a more 'tightly connected' – though not necessarily *larger* - central giant component. Note also the steady decline in both the size of giant component (relative to total node count) and the average component size for both Article and User Networks (Figure 5).

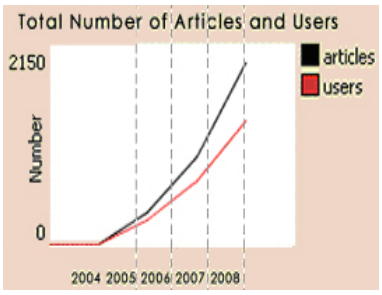

Figure 2. Growth over time of Article and User counts in Cebuano Wikipedia

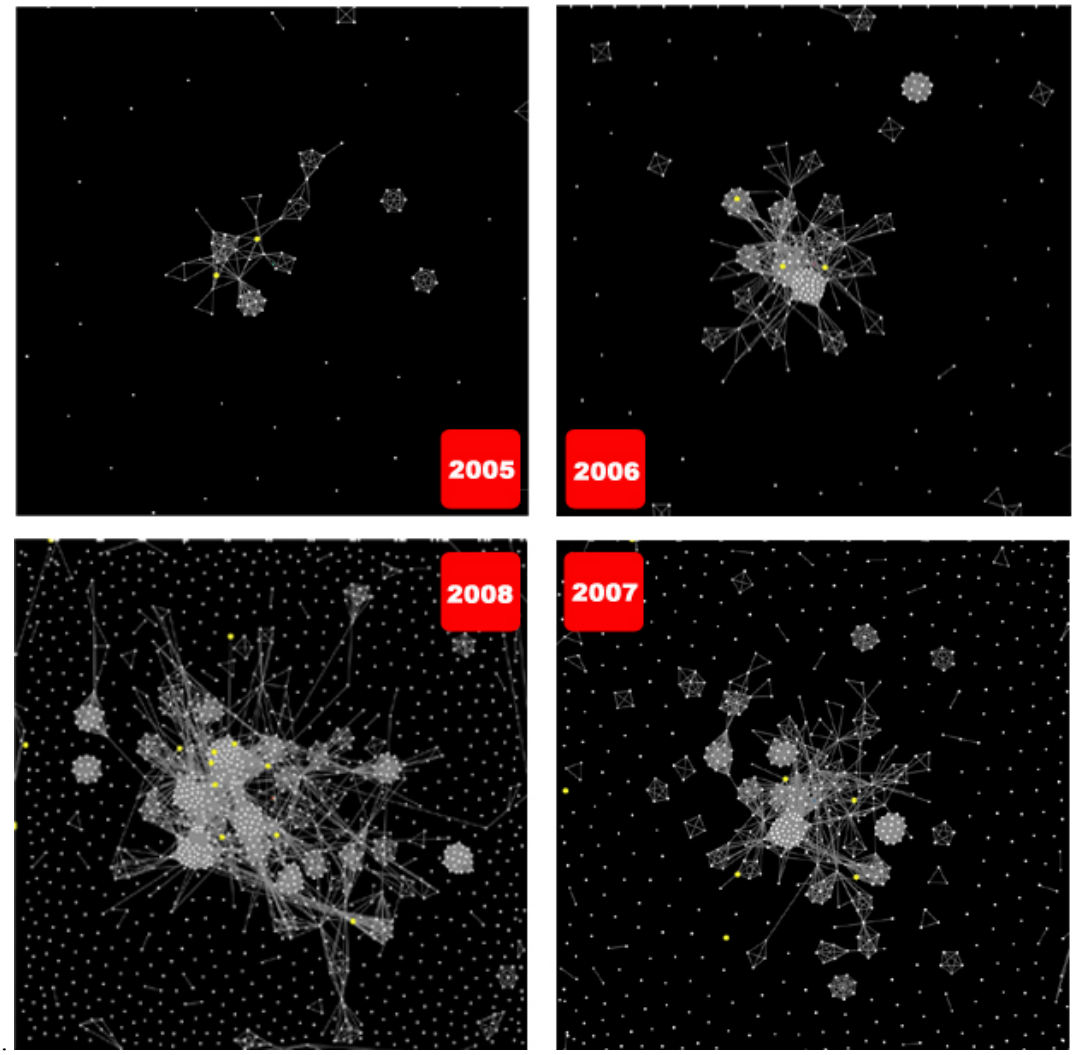

Figure 3. Formation Sequence of Article Network

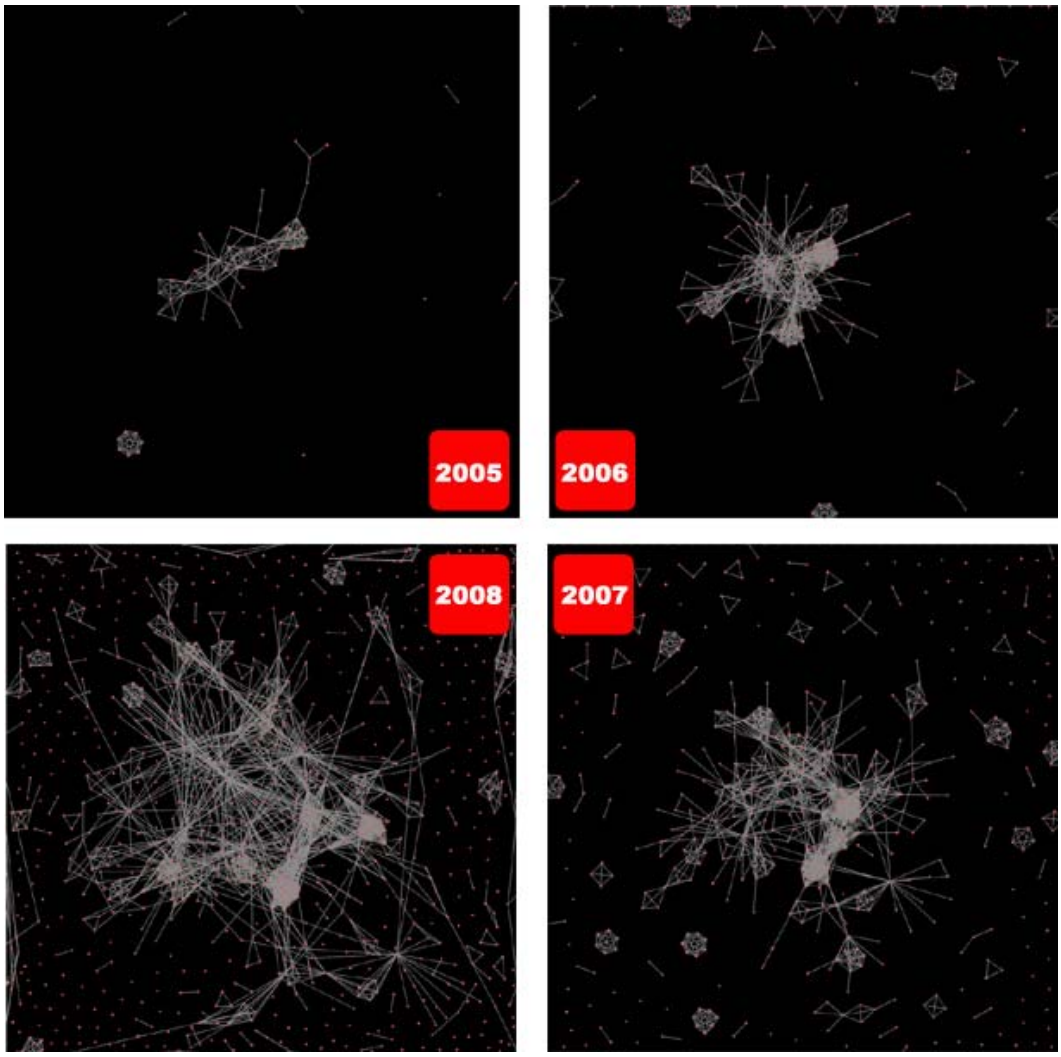

Figure 4. Formation Sequence of User Network

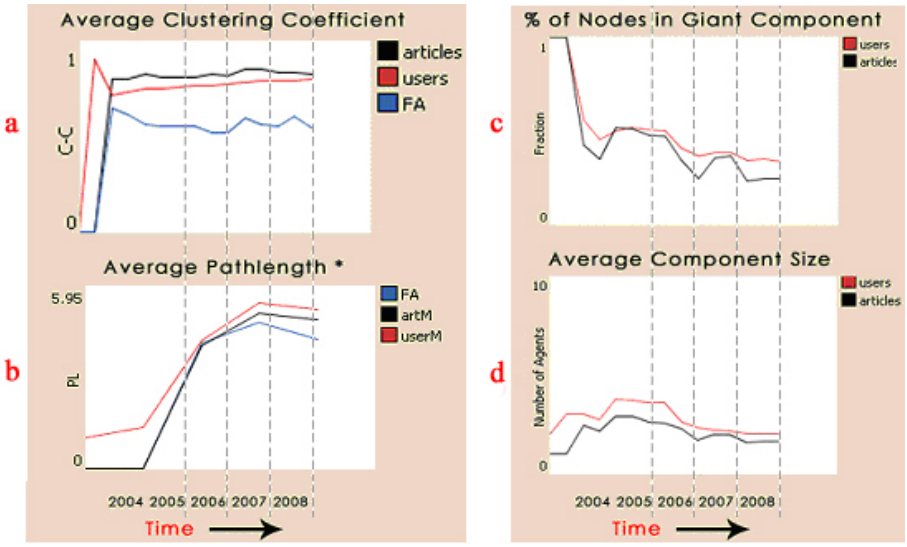

Figure 5. Evolution of structural properties

Small-World Characteristics

The results from Table 1 and 2 demonstrate that Wikipedia's user and article networks are much more clustered than, and have smaller pathlengths than the corresponding sized random networks. We observe that the small world Q of both the article and user networks has been consistently rising (Figure 6) , growing linearly with time, indicating that although the count of users and that of articles have been growing exponentially (Figure 2), the Cebuano Wikipedia network as a whole has been growing 'smaller': knit more tightly together than if it were a random network.

Year (Y)	Clustering Coefficient		Pathlength		Small World Q_a
	Actual	Random	Actual	Random	
2005	**0.79**	0.58	**3.79**	4.71	**1.69**
2006	**0.83**	0.48	**4.16**	5.64	**2.34**
2007	**0.85**	0.50	**5.41**	8.62	**2.71**
2008	**0.88**	0.51	**5.17**	9.93	**3.31**

Table 1. Small World Characteristics of User Network

Year (Y)	Clustering Coefficient		Pathlength		Small World Q_u
	Actual	Random	Actual	Random	
2005	**0.88**	0.64	**4.01**	4.67	**1.60**
2006	**0.89**	0.60	**4.17**	5.81	**2.07**
2007	**0.91**	0.63	**5.06**	8.66	**2.47**
2008	**0.91**	0.66	**4.87**	10.16	**2.88**

Table 2. Small World Characteristics of Article Network

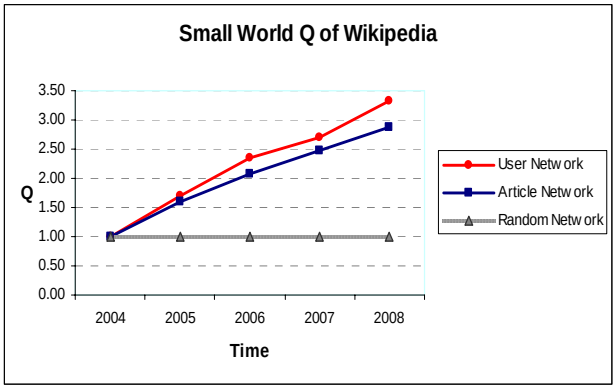

Figure 6. Evolution of Small-World Characteristic Q

Relationship between Network Structure and Quality of Articles

While *Quality* is no doubt a subjective and difficult-to-quantify measure, Wikipedia marks some articles as "Featured Articles". These have been voted by the Wikipedia community to have reached a certain level of encyclopedic excellence (based on criteria such as neutrality, comprehensiveness, referential integrity, style, etc). We are interested in determining if there is any relationship between the *position* of Wikipedia articles in the network structure and their acquiring "Featured" status. Network theory suggests that high quality articles should span structural holes in the article network, thus being *negatively associated* with clustering coefficient and pathlength, as discussed previously.

Structural Property	Average Article	Featured Article Average	Random Network
Clustering Coefficient	0.91	0.61	0.66
Pathlength	4.87	4.22	10.16

Table 3. Featured Articles are have lower than article-average clustering coefficients and pathlengths

Indeed, as Table 3 shows, we find that the average clustering coefficient of featured articles (0.61) is noticeably less than that average clustering coefficient of the network as a whole (0.91) and the average pathlength of featured articles (4.22) is slightly less than the article network average (4.87). This is also clearly visible in Figures 5a and 5b. These results are, however, only indicative. The Cebuano Wikipedia dataset of 1966 articles only contains 10 featured articles, which is a very low sample size. Furthermore, while all featured articles are high quality articles, this does not imply that all high quality articles have already acquired featured article status. Hence, there are difficulties with proposing any test of prediction. While there appears to be a negative relationship of quality (i.e. *Featured Status*) with Clustering Coefficient and Pathlength, there results need to be carefully validated on larger size language Wikipedias. Table 4 provides the results of the logistic regression to determine the relationship between clustering coefficient and pathlength with featured article status.

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
CC	**-3.778**	1.842	4.207	1	.040	.023	.001	.845
PL	**-2.648**	1.671	2.513	1	.113	.071	.003	1.870
Constant	**14.447**	8.074	3.201	1	.074	1880415.716		

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
19.701 (a)	0.331	0.441

a Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The regression equation is **Featured Status = $1/(1+e^{-z})$**

where **z = 14.447 – 3.778 (CC) – 2.648 (PL)**

Table 4. Logistic Regression Results to determine effect of CC and PL on Quality

Degree Correlation

We find that Wikipedia's user network had no significant correlation. However we found strong positive degree correlation in Wikipedia's Article Network. The log-log plot of degree k against the average degree of neighboring nodes (Figure 7) is a straight line that fits the equation $\mathbf{\log(k_{nn}) = 0.58 \log(k) + 0.57}$, with $R^2 = 0.85$**.** This approximates to $\mathbf{k_{nn} \sim k^{0.58}}$. Thus, we find the degree correlation exponent as $\boldsymbol{\beta = 0.58}$, and the Pearson' Correlation Coefficient **r = 0.92**, which is an extremely high positive relationship. To put this in context, r for co-authorship networks in scientific publications within biology and physics disciplines has been found to be 0.127 and 0.363 respectively (Newman, 2002).

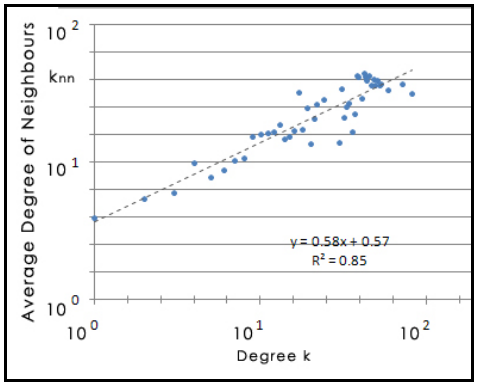

Figure 7. Degree Correlation of Wikipedia's Article Network

Degree Distribution

We find that Wikipedia's interaction networks have a scale free degree distribution. The log-log plots of Figure 8 are straight lines, confirming power laws. This implies that in both Article and User networks, there are a large number of nodes with very few links, and a few nodes with a large number of links. We see that the power law exponent γ is increasing over time (Table 5), indicating that the in-egalitarian nature of the 'rich get richer' scale-free degree distribution is increasing.

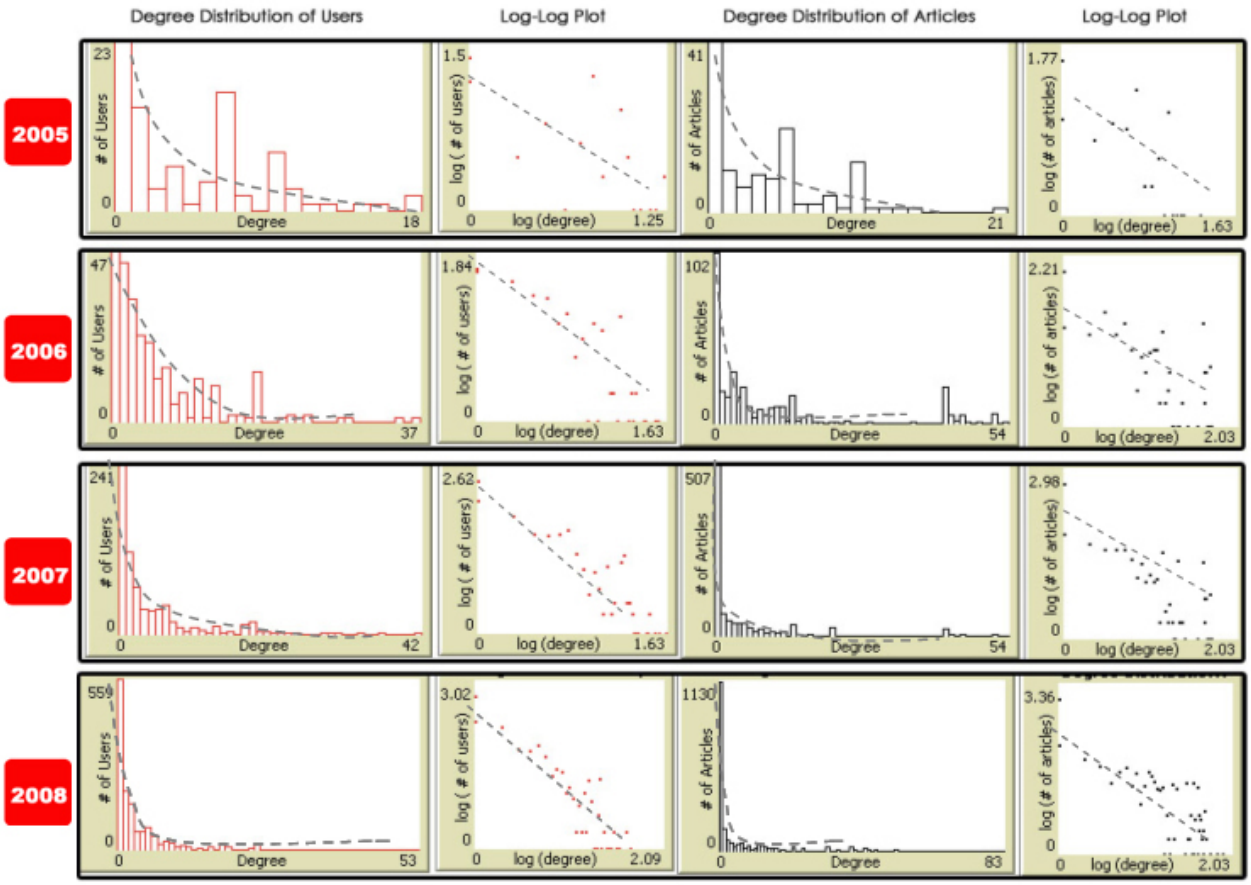

Figure 8. Evolution of Degree Distribution (showing log log plot) for Cebuano Wikipedia User and Article Networks

Year	Power Law [$P(k) \sim k^{-\gamma}$] : Exponent Estimation					
	Wikipedia's User Network			Wikipedia's Article Network		
	γ_u	R^2	C.I *	γ_a	R^2	C.I *
2005	**0.81**	0.47	0.29	**1.07**	0.52	0.28
2006	**1.20**	0.73	0.18	**1.01**	0.69	0.17
2007	**1.50**	0.84	0.15	**1.14**	0.72	0.19
2008	**1.64**	0.85	0.14	**1.25**	0.78	0.16

Table 5. Evolution of Power Law Exponent γ in Cebuano Wikipedia's User and Article Networks (*α =.12)

INTERPRETATION AND DISCUSSION

We have found a definite trend of **increasing small world characteristic Q** of Wikipedia. We observe from Table 1, 2 that this is mainly due to pathlength not increasing at the rate corresponding to the random graph. One may attribute this to the increasing scale-free nature of degree distribution - The power law exponent γ has been steadily rising over time. The rising influence of *hubs* – high degree nodes - on connectivity of the network keeps average pathlength low. A possible cause of the increasing power law exponent γ is the high *degree correlation* of the article network, which implies that high degree articles preferentially link to other high degree articles.– this is a direct consequence of the often cited "edits beget edits" rule: An article of high current relevance receives disproportionately high user attention, which translates to user edits, which translate to more edits (empirically shown also by Buriol et al, 2006). The consequences are an increasingly in-egalitarian distribution of links in the article network – a scale free distribution , where 'the rich get richer', and become the 'hubs' of the network. They are, by their nature, centrally located and connect many shortest-paths across the network. Since the clusters they inter-connect were hitherto unconnected, the hubs span structural holes, and also, by the same token, have lower pathlengths than the network average. The hubs are in a unique position of a) Having access to non-redundant sources of contributions and b) Having access to a disproportionately large quantity of contributions. A combination of (a) and (b) results in the preferential emergence of Featured Articles at the hubs. We believe this is the **mechanism by which structural holes come to be associated with Quality**.

Another important aspect is **growth**. Figure 2 shows that the number of users and articles is growing exponentially. As Figures 3 and 4 show, initially, both networks contain one main component, with many small disconnected clusters. As hubs emerge (by the structural hole spanning mechanism described above), they serve as connectors between clusters. Guimera et al (2005), studying the network growth of scientific co-authorships, observed a similar process, of cluster "overlapping". In their study, the co-authorship network became denser and eventually, all the nodes were connected together in one giant component (The formation of 'the invisible college'). In Wikipedia, while the cluster overlap process is going on, new users and articles are continually joining the network. This joining rate is much higher in Wikipedia than in the scientific co-authorship network (where the rate was linear). Since the growth of new users and articles is exponential, even as the giant component in absolute terms expands, it decreases in percentage terms (Figure 5c). Many of the new users and articles form new clusters outside the giant component, and again, a few hubs emerge that cross-link clusters. This way, the process repeats itself and wikipedia grows in size. Amidst this 'self-similar' process, it is interesting to note that there is no 'tipping point' (unlike in many other studies of percolation).

Our results emphasize the importance of the hubs in both the user and article networks. To place things in context, the hubs in our user network would be equivalent to Kittur et al's (2006) 'Elite Users' and Anthony et al's (2005) 'Zealots' – dedicated super-editors who numerically make up most of the contributions on Wikipedia, and 'hold the small-world together'. Also, an important finding from Wikipedia research - that 'multiple diverse viewpoints' are responsible for quality - is supported by our result that Featured Articles emerge from hubs spanning structural holes - hubs have high degree (a large number of unique contributors) and low clustering coefficient (they connect otherwise unlinked 'diverse' clusters). Our finding that '**hubs hold the growing small world together'** has an important implication for practitioners: Organizations should place an emphasis on identifying and nurturing this class of users and articles in the network, from an early stage. This emphasis may be embedded in the design of the collaborative knowledge system itself, in the form of expert features, as well as provision of powerful policy tools to enable and empower hubs.

LIMITATIONS

Two users with a large number of articles in common may have a 'stronger tie' than if they had only one article in common: Our current approach does not focus on this *'strength of ties'*. Secondly, we need empirical results from other, larger language Wikipedias to ascertain the network association of quality. Results obtained from Cebuano Wikipedia at individual featured article level are strongly indicative, but need to be confirmed with other language Wikipedias. Thirdly, the interaction data is imperfect : Not all interactions are the same. For e.g. a contribution of one word is probably qualitatively different from a contribution of three paragraphs of text, but there is no distinction in the meta-revision history dataset we use.

FUTURE WORK

A first step forward from here would be validation of these results on larger language Wikipedias, and investigation of the effects of "*strength of ties*". We believe there are still many things to learn about collaborative knowledge creation from Wikipedia, and we are hopeful this work will provide a useful beginning. One possible direction of research is to investigate the interplay between social norms, interaction network structure, technology architecture and governance in open collaborative systems. In Wikipedia, for instance, social norms play a major role in governance, and consequently, in quality of content. *Persistence* of (*often beneficial*) early social norms, even in the face of massive user-base churn, is a puzzle that could be solved using our findings, in conjunction with results from epidemiology, that suggest that in *specific* types of interaction networks, with dense inner cores, an 'infection' could survive indefinitely. This would require modeling norms as *memes* – socially transmissible ideas. Another perspective is the well-documented robustness characteristics of scale-free networks, which have been found to be extremely resilient to random failure i.e. removing a node at random has a very low probability of affecting the connectivity of the network. An explanation for Wikipedia's success based on the robustness analysis of Wikipedia's scale free user network could be another fruitful line of enquiry. An interesting application would be a second order *predictive model* of open collaboration, using Agent-Based Simulation. This would be of interest for businesses using collaborative technology either *internally*, for knowledge management, or *externally*, for crowd-sourcing and engaging with consumers. In a broader context, an extension of this approach to studies at multiple scales of analysis (individual, group, organizational) may go some way in answering the call in organizational literature for a unifying framework for knowledge networks.

REFERENCES

1. Anthony, D., Smith, S., and Williamson, T. (2005) Explaining quality in Internet collective goods : Zealots and Good Samaritans in the case of Wikipedia. http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf. (July 3, 2007)
2. Barabási, A. and Albert, R. (1999) Emergence of scaling in random networks, *Science*
3. Benkler, Y. (2006) The wealth of networks: how social production transforms markets and freedom. Yale University Press, New Haven.
4. Borgatti, S. and Pacey, C., (2003) The network paradigm in organizational research: A review and typology. *Journal of Management*, 29: 991–1013.
5. Buriol, L., Castillo, C., Donato, D., Leonardi, S. and Millozzi, S. (2006) Temporal analysis of the Wikigraph. *In Web Intelligence Conference.*
6. Burt, R. (1992) Structural Holes: The social structure of competition. Cambridge, MA: Harvard University Press.
7. Burt, R. (2005) Brokerage and closure: An introduction to social capital. New York, NY: Oxford University Press.
8. Chesney, T. (2006) An empirical examination of Wikipedia's credibility. *First Monday*, volume 11, number 11
9. Coleman, J. (1990) Foundations of social theory. Cambridge, MA: Harvard University Press.
10. Cowan, R. (2004) Network models of innovation and knowledge diffusion. *MERIT* - Infonomicus Research Memorandum Series.
11. Davis, G., Yoo, M. and Baker, W. (2003) The small world of the American corporate elite: 1982–2001. *Strategic Organization*, 3: 301–326.
12. Dutta, A., Roy, R. and Seetharaman, P. (2008) Wikipedia usage patterns: The dynamics of growth, *Proceedings of the International Conference on Information Systems* , Paris.
13. Erdos, P. and Renyi, P. 1959. On Random Graphs. Publ. Math. Debrecen.
14. Guimerà, R., Uzzi, B., Spiro, J. and Amaral, L. (2005) Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance, *Science*.
15. Hafner, K. (2006) Growing Wikipedia refines its 'anyone can edit' policy, *New York Times*
16. Kakihara, M. and Sorensen, C. (2002) Exploring Knowledge Emergence: From Chaos to Organizational Knowledge. *Journal of Global Information Technology Management.* Vol.5, No.3, pp. 48-66.

17. Kittur, A., Chi, E., Pendleton, B., Suh, B. and Mytkowicz, T. (2007) Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *Annual ACM Conference on Human Factors in Computing Systems*.
18. Milgram, S. (1967) The small world problem. *Psychology Today*, 2,60-67.
19. Newman, M. (2002) Assortative mixing in networks. *Physical Review Letters*
20. Orlikowski, W. (1992) The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, 3(3): 398-427.
21. Parameswaran, M. and Whinston, A. (2007) Research issues in social computing, *Journal of AIS*, Volume 8, Issue 6, Article 1, pp. 336-350.
22. Schilling, M. and Phelps, C. (2004) Small world networks and knowledge creation: Implications for multiple levels of analysis, *Academy of Management Conference*, New Orleans
23. Stacey, R. (2000): The emergence of knowledge in organizations. *Emergence*, vol. 2, no. 4, pp.23-39.
24. Uzzi, B. and Spiro, J. (2005) Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111: 447–504.
25. Wagner, C. and Majchrzak, A. (2007). Enabling customer-centricity using wikis and the wiki way. *J. Manage. Inf. Syst.,* 23 :17-43.
26. Wasserman, S. and Faust, K. 1994. Social network analysis: Methods and applications. New York, NY & Cambridge, UK: Cambridge University Press.
27. Watts, D. (1999) Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton, Princeton University Press.
28. Watts, D. and Strogatz, S. (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440-42.
29. Tisue, S. and Wilensky, U. (2004) NetLogo: A Simple Environment for Modeling Complexity. *Proceedings of the International Conference on Complex Systems*.