Can We Data Mine Open Source Software Development?

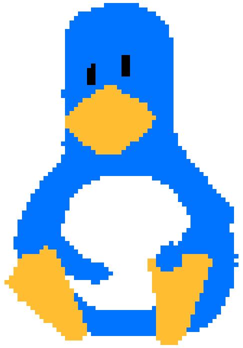
George Kuk

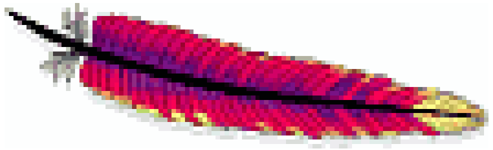
Nottingham University Business School

Agenda

- Context and research in OSS development
- Empirical findings in relation to knowledge sharing and innovation
- Limits and myths of data mining knowledge (generation and renewal) in electronic networks of practice

OSS: Creatures Big and Small





K Desktop Environment

mozilla.org

Hallmarks: Communities of Practice

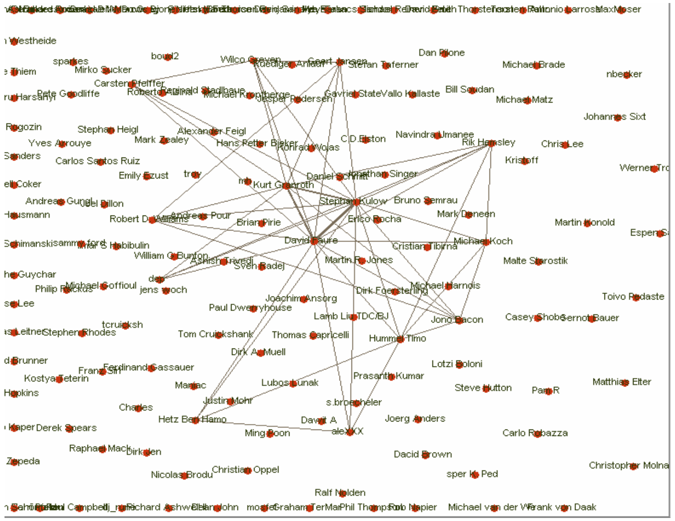
- According to Wenger (1998, pp.125–126)
- Epistemic Parameters
  - □ (1) Sustained mutual relationships—harmonious or conflictual
  - □ (2) Shared ways of engaging in doing things together
  - □ (3) The rapid flow of information and propagation of innovation
  - □ (4) Absence of introductory preambles, as if conversations and interactions were merely the continuation of an ongoing process
  - □ (5) Very quick setup of a problem to be discussed
  - □ (6) Substantial overlap in participants' description of who belongs
  - □ (7) Knowing what others know, what they can do, and how they can contribute to an enterprise
  - □ (8) Mutually defining identities
  - □ (9) The ability to assess the appropriateness of actions and products
- Structural Parameters
  - □ (10) Specific tools, representations, and other artifacts
  - □ (11) Local lore, shared stories, inside jokes, knowing laughter
  - □ (12) Jargon and shortcuts to communication as well as the ease of producing new ones
  - □ (13) Certain styles recognized as displaying membership
  - □ (14) Shared discourse reflecting a certain perspective on the world

OSS Dynamics: The 3 Cs Model

Coordination, Concentration and Configuration

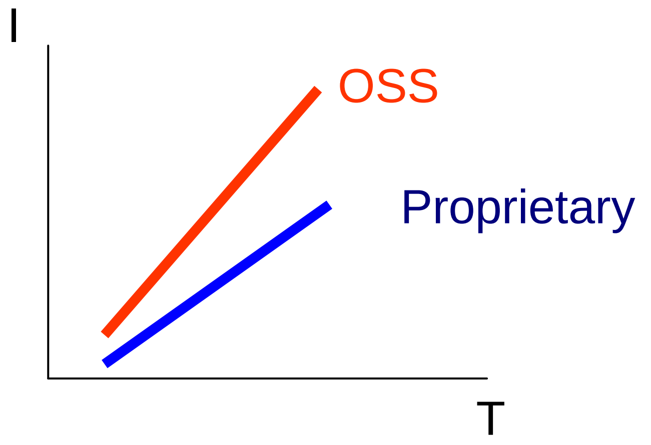
OSS Dynamics: 3 Cs

Coordination	1 to Many n to N, where n ⊂ N	Balance between coordination overheads and network heterogeneity
Concentration	$p_k \sim 0.832\, k^{-3}$ $p_k \sim 12\, k^{-3}$	Balance between exploration and exploitation; critical mass under strategic interaction between 2.3% (random) to 33% (preferential attachment)
Configuration	2-tier structure	Organize sufficiently to generate knowledge and problem-solving capability



Coordination: Findings

Innovation is linear and multiple, achieves a faster rate when compared to proprietary software development



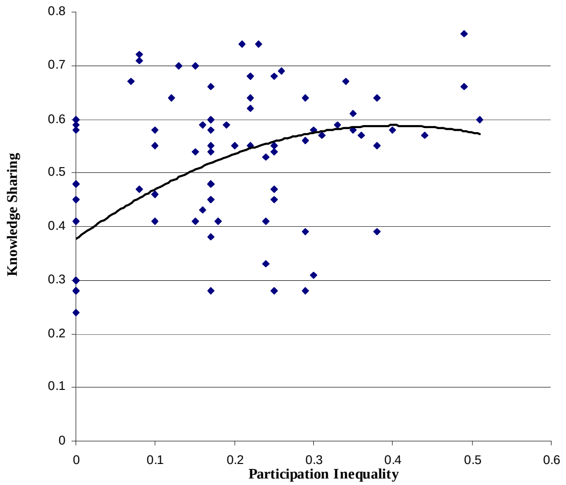
Note: Software codes and bugs between minor and major OSS releases were data mined

In contrast to proprietary model, by democratizing innovation, the OSS communities have to rely heavily on feedback loops with single, unitary positive feedback loops to increase individual contributions; and double feedback loops to mitigate too much diversity and produce integrated products

Concentration: Findings

Developers strategically interact with others to concentrate exploitation at the core and increase exploration at the periphery

Figure 1. The Curvilinear Relationship between Participation Inequality and Knowledge Sharing represented by Y = -1.3173X^2 + 1.056X + 0.3768 and R-squared = 0.34.



The rate of innovation (knowledge sharing) exhibits a curvilinear relationship with concentration, i.e. too little or too much concentration is bad for OSS development and innovation

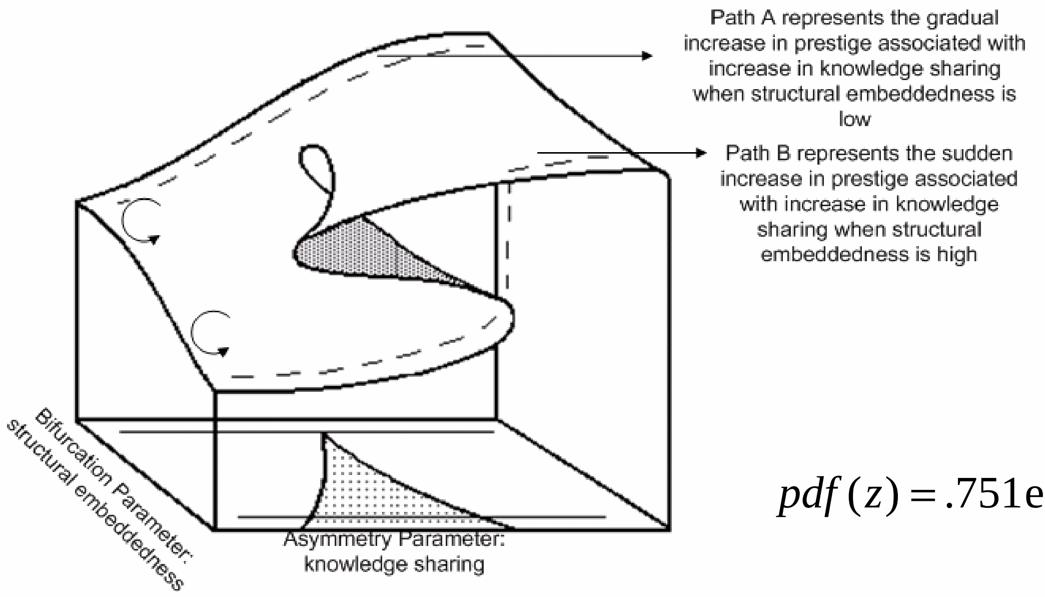
Knowledge sharing based on interpretative analysis (a form of human instantiation);

Even with human coders, the inter-coder reliability Kappa drops drastically if information is scrambled but increases after considering the ebb and flow of epistemic exchanges

Configuration: Findings

The cusp catastrophe model can be used to model the emergence of the 2-tier structure

It shows the complex interplay between social interconnectivity (a structural parameter) and knowledge sharing (a epistemic parameter)



Path A represents the gradual increase in prestige associated with increase in knowledge sharing when structural embeddedness is low

Path B represents the sudden increase in prestige associated with increase in knowledge sharing when structural embeddedness is high

$$pdf(z) = .751\exp[-.016z^4 + .216z^3 - .763z^2 + .830z]$$

$$pdf(z) = .707\exp[-.027z^4 + .385z^3 - .213bz^2 + .458az]$$

$$pdf\ (z) = .555\ \exp[.\,018\,z^4 - .149\,bz^2 + .414\,az\,]$$

Can we data mine OSS Developments?

Coordination

Using OSS communities as an unit of analysis ✓

Restricted to codes and bugs

Concentration

Knowledge types (latent semantic analysis) ✓

But for knowledge sharing considering the knowing aspects of knowledge and contextual information ✗

Configuration

Consider the importance of structural and epistemic parameters in our understanding of the emergence of the 2-tier structure ✗

Limits of Data mining K

- Data mining as an inductive tool to understand the back box of knowledge sharing can exacerbate the danger of “garbage in and garbage out” by adding “garbage throughout”
- Data mining needs bounded rationality involving human instantiation (parameterization) and heavy interpretation (theorizing)
- It requires damage limitation/controls due to the decontextualization of the tacit element of knowledge and knowing; and the deconstruction of the integral relationship between structural and epistemic parameters in our understanding of emergence of new forms of learning/organization structures and notably knowledge creation