A “Social Capital Perspective” of Participant Contribution in Open Source Communities - The Case of Linux

Myong Rae (Ray) Chang, Ph.D.

July 14, 2011
OSS Development Process: Network Capital → Contribution

Agenda

- Motivation
- Theoretical Challenges
- Research Model
  - Four Dimensions of Network Capital
- Method
  - Network Measures
  - Content Analysis
- Results
- Analyses Implications
  - Limitations & Future Study

LINUX & OSSD

Thread-level Dynamic Process

PLS
**The Linux Kernel**

- Desktop/Laptop: **3%**, Server/Mainframe: **65%**, Supercomputer: **92%**
- Smartphone (Android): **25%**  
  [rough averages from multiple sources, worldwide, 2010 ]

*Linus Tovalds*, Jan. 2011

“Not just Android. What I’ve found that has been most fun for me has been when people are using Linux in ways that I don’t use it or in ways that I never intended it to be used, people using it in embedded areas, and with cellphones like Android but also all the crazy people using it in printers and TVs.”
Linux Open Source Community

- The largest OSS development community
- Being evolved dynamically at every second
- A “virtual” workplace open to any participants from any place in the world at any time (but only “hackers” survive)
- “Open” to any contribution on “voluntary” basis (note: many are now paid workers from Linux-related companies)
- All “peer-reviewed” process: from ideas to codes & feedbacks
- Driven by crowd wisdom, not by dedicated plans or profits
- Highly technical and rigorous discussions: it should work!
- Administered by “Maintainers” of numerous subsystems
✓ A history of intellectual knowledge-exchanges on a *single* specific topic

http://www.gossamer-threads.com/lists/linux/kernel/655933
A Message-exchange “Network” of a Thread
= a knowledge exchange network built on reply-to relations formed by the messages within a thread

Node: each participant
Link or tie: relation of in-reply-to (with the attribute of message multiplicity)

Note: Although each message is “broadcast” to all subscribers of the mailing list, only respondents to the message would take special meaning w.r.t. the topic; i.e., the network is a construction based on “reciprocity” relationship, a characteristics of online relationship.
Patterns of Thread Network

A “signature” of how they have communicated/collaborated on a specific task.
Evolution of A Thread Network

Thread# 913809

Network Size vs. % Messages (Time)

seed
Dynamics of Network Measures

![Graph showing the dynamics of network measures](image-url)

- **Density**
- **Avg. Geodesic Distance**
- **Degree Centralization**
- **Betweenness Centralization**

Graph-axis labels:
- **Relative Change of Network Measure**
- **% Messages**

The graph illustrates the relative change of various network measures as a function of the percentage of messages transmitted.
We are interested in “early-stage” network building to influence the thread performance in the later stages.
Research Questions

- Will the early-stage accumulation of network capital affect participant contribution in the later stage?

- Then, what types of network capital will be associated with participant contribution in terms of quantity & quality?

- During the lifecycle of OSSD, how can we help to elicit “more” and “better” participant contribution?
Theoretical Background

- Social Capital Perspectives in Online Communities
  - Structural properties (position, structure)
    → individual’s knowledge sharing behavior
    [Cross & Cummings 2004; Nerkar & Paruchuri, 2005]
  - Role of relational properties → learning and knowledge transfer
    [Hansen 1999; Uzzi & Lancaster 2003]
  - Various types of social capital (structural, cognitive, relational)
    → knowledge sharing of individuals and groups
    [Wasko & Fajar 2005 in electronic networks of practice; Kuk 2006 in an OSS community]

- Theoretical Challenges
  - Fragmented view vs. a “comprehensive” framework: a holistic model integrating various dimensions of social capital
  - Extant isolated element-level approach (actor, dyadic link, ego-centric [Borgatti & Foster 2003]) vs. understanding of actors’ “collective behavior”: a network (e.g., thread) level approach
  - “Static” snapshot approach vs. “Dynamic” aspect of social capital
## Four Dimensions of Network Capital

<table>
<thead>
<tr>
<th>Dimension</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network</strong></td>
<td><strong>Structural Capital</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Network centralization</td>
<td></td>
</tr>
<tr>
<td><strong>Link</strong></td>
<td><strong>Relational Capital</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Network strength</td>
<td></td>
</tr>
<tr>
<td><strong>Node</strong></td>
<td><strong>Governance Capital</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Administrator participation</td>
<td></td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td><strong>Dynamic Capital</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Network growing speed</td>
<td></td>
</tr>
</tbody>
</table>

- Time: 6 hrs, 12 hrs
Research Model: PLS

Growing-stage Network Capital (1/3)

Network Centralization

Network Strength

Administrator Participation

Network Growing Speed

OSS Performance (2/3)

Contribution Quantity

Code in Initial Message

Inhibiting Climate

H1a

H1b

H2a

H2b

H3a

H3b

H4a

H4b

H5

positive

negative
Hypotheses 1a/1b

Two Contrasting Views on Centralization

- Discourages diverse views
- Negative impact on creativity
- Reduces autonomy of participants

vs.

- Rapid diffusion of innovative knowledge
- Easy access to experts with lower cost
- Integration of diverse ideas

In Thread-level Collaboration:

- Benefits outweigh the costs from the lack of idea diversity
- Hub-structure facilitates member contribution
- Continuous review and feedback systems
- Multiplicity of views and ideas more easily integrated
**Hypotheses 2a/2b**

**Relational Capital**

Network Strength

- **H2a**
- **H2b**

**Contribution**

- **Quantity**
- **Quality**

**✓ Tie Strength (link thickness)**

- Useful conduits for knowledge exchange (in many social network studies)
- Essential for substantive contribution

**✓ Network-level Strength**

- A network is stronger if containing more strong ties.
- Building a normative environment fostering collaboration and coordination
- A sense of “reciprocity” ensures continuing supportive exchanges and generating in-depth discussions.
Hypotheses 3a/3b

- **In Traditional Organization Settings**
  - Leader’s involvement is effective for affective and continuance contributions.
  - Strong governance encourage self-concept-based motivation.

- **In OSS Communities**
  - Admin’s tight control evokes a cathedral type of decision making structure.
  - Undermining participants’ autonomy and sense of ownership [von Krogh 2003].
  - Members do not participate when most work is conducted by a leader.
Research on Interpersonal Communications

- The quantity and quality of exchanged knowledge are highly associated with the rate at which knowledge is delivered [Carlson & Zmud 1999].
- Faster responses allow receivers to act upon in a timely manner.

In Knowledge-intensive Online Communities

- Level of detail or extensiveness is often more important than response time.
- Rapid responses might decrease the perceived value of knowledge exchanged.
- Slow responses tend to be more rational and cognitive (faster ones be more emotional).
- Developers place more weight on “accuracy”: slow responses reduce uncertainty and equivocality by providing more complete knowledge [Weiss et al. 2006].


**Hypotheses 5 and Controls**

- **Quantity → Quality**
  - Increasing # of proposed ideas
  - → higher probability of quality ideas
  - [Barki & Pinsonneault 2001]

- **Controls**
  - Initial message with codes proposes more concrete ideas to call for more contribution [Roberts et al. 2006]
  - Inhibiting culture structurally prevent members’ contribution [Bogozzi & Dholakia 2006]
RFC (Request For Comments) threads collected from archive
- Suggestion/discussion/collaboration on new innovative ideas & features
  (over 90% of them had less than 15 messages)
- Used threads with enough(>25) messages → **223 RFC threads**
- For each thread, a matching set of a network file and a text file of messages were created (only original contents used after removing quotes, program codes, etc.)

**LINUX Community**

**223 Networks** (growing stage)

**Computing Measures**

**OSS Contribution**

**LIWC**

**Network Measures**

**C/C++**

**223 RFC Threads**

**223 text files of whole messages**

**C/C++**
## Operationalizing Network Capital

### Growing Stage* of a Thread

<table>
<thead>
<tr>
<th>Sample Period</th>
<th>Construct</th>
<th>Item</th>
<th>Definition</th>
</tr>
</thead>
</table>
| Growing Stage* (Initial 1/3) of a Thread | Network Centralization | degree centrality | $X_i$: degree/betweenness centrality of node $i$  
$X^*$: maximum degree/betweenness centrality |
|                          |                                | betweenness centrality      | Centralization = $\frac{\sum_i (X^* - X_i)}{\text{Max} \sum_i (X^* - X_i)}$ |
|                          | Network Strength               | multi-link ratio            | proportion of strong ties in a thread = # strong ties / # total links |
|                          |                                | multi-message ratio         | message overloading per link in a thread = # total messages / # total links |
|                          | Administrator Participation    | admin-node ratio            | proportion of administrators in a thread = # administrators / # total participants |
|                          |                                | admin-message ratio         | proportion of administrators’ messages = # messages of admins / # total messages |
|                          | Network Growing Speed          | message entering rate       | log (message occurring rate per unit time) = log (# total messages / total elapsed hours) |

* Note: for capturing the growing-stage dynamics, during the sample period we measured each item three times with the increase of message volume and used averaged values in the analysis.
Adopting Objective Content Analysis

- An automated text analysis tool, LIWC, that calculates the degree to which people use different categories of words
- Internal dictionary: 4,500 words and word stems, 32 word categories (average word-capture percentage: over 86%)
- Categories: cognitive, positive/negative emotions, insight, causation, etc.
- Text content → word count or percentage by category
- Broadly used in social sciences, linguistics, health, information science, etc.

Contribution Quality: percentage of words in category “INSIGHT”

- Words associated with “learning” and “understanding” [J. Pennebaker et al.]
- Actively thinking about something in a self-reflective or insightful way
- Generally a class of verbs called ontological verbs and included words: understand, realize, know, meaning, etc.
- Health and immune function [Klein & Boals, 2001; Petrie et al. 1998]
- Experiential learning [Abe 2009]

Contribution Quantity: word count per message in a thread
## Operationalizing Contribution & Controls

### Dependent Variables

<table>
<thead>
<tr>
<th>Sample Period</th>
<th>Construct</th>
<th>Item</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable and Mature Stage (Later 2/3) of a Thread</td>
<td>Contribution Quantity</td>
<td>word count per message</td>
<td>log (word count per message for a thread)</td>
</tr>
<tr>
<td></td>
<td>Contribution Quality</td>
<td>Insight word percentage</td>
<td>percentage of words in a thread which belong to LIWC category “Insight”</td>
</tr>
</tbody>
</table>

### Control Variables

<table>
<thead>
<tr>
<th>Growing Stage (Initial 1/3) of a Thread</th>
<th>Code in Initial Message</th>
<th>ini-code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Code in Initial Message</td>
<td>ini-code</td>
<td>1 if program codes are found in the thread-initiating message 0 otherwise</td>
</tr>
<tr>
<td></td>
<td>Inhibiting Climate</td>
<td>Inhibit word percentage</td>
<td>percentage of words in a thread which belong to LIWC category “Inhibit”</td>
</tr>
</tbody>
</table>
Results: Structural Model

Network Centralization

Network Strength

Administrator Participation

Network Growing Speed

Contribution Quantity

R² = 0.120

Contribution Quality

R² = 0.085

Code in Initial Message

Inhibiting Climate

Notes: t-statistics are in parentheses. +p<0.1; *p<0.05; **p<0.01
Supported Hypotheses

- **Centralized** communication structures can serve as effective conduits for quality discussions only when “naturally” established by peer members; administrators’ “artificial” interventions/controls would result in adverse outcomes (H1, H3)
- **Administrators** need to keep their intervention to a minimum during the growing stage of a thread; non-admin peer participants with relevant expertise should lead the discussion (H1, H3)
- For more contribution, participants need fully utilize pre-established communication relationships (H2) and **take time** in responses to provide accurate and detailed information (H4)
Unsupported Hypotheses

- **H1a**: Mixed effects from two opposing forces (diversity vs. learning)
- **H2b**: Strong ties may produce many redundancies, preventing the generation of innovative ideas  
- **H4b**: Contribution quality does not automatically occur with the progression of time
- **H5**: Voluminous discussions are not essential for qualitative contributions
Limitations and Future Research

Limitations
- In staging the OSSD lifecycle, the approach using real time-clock needs to be employed along with message volume approach
- Automated text analysis and objective assessment still need more validation to be used in IS context
- Limited variance explanation suggests the existence of unobserved factors affecting the members’ contribution

Future Research
- Abundant research challenges exist to advance the understanding of the evolutionary aspects of OSSD collaboration in network/thread level
- Exploring the changes in the structural formations with the new entry of participants; e.g., with whom do the new entrants tend to interact? Among the new entrants, who contributes the most in terms of network capital and participant contribution and why?
- Network simulation models for the dynamic process of online collaboration and social media activities (e.g., on Facebook, Twitter)