MSR: Mining for Scientific Results?

Jim Herbsleb
School of Computer Science
Carnegie Mellon University
jdh@cs.cmu.edu
http://conway.isri.cmu.edu/~jdh/

The author gratefully acknowledge support by the National Science Foundation under Grants IIS-11 0414698, IIS-0534656, OCI-0943168, and IGERT 9972762, as well as the Software Industry Center at CMU and its sponsors, particularly the Alfred P. Sloan Foundation.
MSR and the Value of Prediction

- High impact relative to most SE research
- Practical utility
- Goal is prediction – Insight and understanding are optional
NEW CUYAMA

Population 562
Ft. above sea level 2150
Established 1951
TOTAL 4663
MSR 2010 Topics

- Predicting
  - Bug severity
  - Number of bugs (2)
  - Fault-proneness
  - Efficiency
  - Change

- Comparing
  - Precision finding bugs
  - Using stack traces

- Detecting
  - Security bugs (2)
  - Clones (3)
  - Metapatterns
  - Licenses
  - Occasions to contribute

- Modeling evolution
- Methods (7)
- Others (4)
Since MSR Is So Successful . . .

- Why might you want to do something a bit different?
- What is it exactly that I’m suggesting some of you might wish to do?
To Bleed or not to Bleed . . .

- Late 18th century
- Francois Joseph Victor Broussais
  - Chief physician Paris military hospital
  - Promoted bleeding of “affected organ”
- Pierre-Charles-Alexandre Louis
  - Actual data collection about outcomes
  - Bleeding is not such a great idea
Mining Medical Repositories (MMR 1780)

- Predicting
  - Severity
  - Who will become ill
  - Changes in condition
- Comparing
  - Treatments
  - Physicians
  - Hospitals
- Detecting
  - Presence of a disease
  - Type of injury
  - Patterns of outbreaks
Statistics, Medicine, Science

- Pierre Louis promoted use of correlation of treatment and outcome to evaluate effectiveness
- Others, e.g., Friedrich Oesterlen, denied that this was science
  - Discovery of correlation not science
  - Science requires understanding the causal connection
- Joseph Lister – outcomes of antiseptic surgery in Edinburgh
  - Mortality rates decreased from 45.7% to 15%
  - Technique based on Louis Pasteur’s “germ theory”

Source: Chen, T.T., *History of Statistical Thinking in Medicine*
The Scientific Method?

- Paul Feyerabend
  - “Anything goes!”
- Argues that methods grounded in particulars of each science
  - Questions they ask
  - Phenomena they study
- All agree that theory is central
- “Scientific theory is a contrived foothold in the chaos of living phenomena.”
  - Wilhelm Reich
A Definitive Review of Relevant Scientific Theories
An Idiosyncratic Selection of Two Possibly Relevant Theories I Happen to Have Heard of . . .

- Based on a stylized narrative that predicts statistical associations among variables
Social Psychology Theory:
Collective Effort Model

Social Network Theory: Knowledge Transfer

<table>
<thead>
<tr>
<th>TIE STRENGTH</th>
<th>KNOWLEDGE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strong</td>
<td>Weak</td>
</tr>
<tr>
<td>Noncodified,</td>
<td>Low search benefits, moderate transfer problems</td>
<td>Search benefits, severe transfer problems</td>
</tr>
<tr>
<td>Dependent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codified,</td>
<td>Low search benefits, few transfer problems</td>
<td>Search benefits, few transfer problems</td>
</tr>
<tr>
<td>Independent</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Theorizing about Coordination

• Collaborators
  • Beki Grinter
  • Audris Mockus
  • Marcelo Cataldo
  • Patrick Wagstrom
  • Kathleen Carley
  • Laura Dabbish
  • Anita Sarma
Conway’s Law

• “Any organization that designs a system will inevitably produce a design whose structure is a copy of the organization's communication structure.”*

• Modularity is an effective coordination strategy

• Product modularity leads to work modularity, which structures organizations**

---


Conway’s Law

Components

Teams

Software

Organization

Isomorphism
Conway’s Law

Components

Teams

Homomorphism

Software

Organization
Modularity: Just a Good Start

- Modularity is never perfect -- how can we characterize intermediate states?
- Teams and modules are constantly changing . . .
- How does work become coupled?
- What does coupling of the product imply about how the people do the work?
What would a good theory look like?
Coordination and the Kinetic Theory of Gases
Software Development

- people
- time

- Development Work
- Decisions
- Constraints
Key Definitions - 1

Project is a set of engineering decisions $X_i$

Feasibility function:
$$f\left(x_{1j(1)}, \ldots, x_{nj(n)}\right) = \begin{cases} 1 \text{ iff product satisfies requirements,} \\ 0 \text{ otherwise} \end{cases}$$

Feasible choices, $FC(X_i)$, is the set
$$x_{ij*} : \forall k \neq i, \exists j(k) \text{ such that}$$
$$f\left(x_{1j(1)}, \ldots, x_{ij*}, \ldots, x_{nj(n)}\right) = 1$$

Key Definitions - 2

Effects of a decision:
\[ j(k) : X_k = x_{kj(k)} \text{ on a decision } l \]
\[ E(X_l \mid X_k = x_{kj(k)}) \text{ is the set difference} \]
\[ FC(X_l) - FC(X_l \mid X_k = x_{kj(k)}) \]

Maximal effects of a decision:
\[ ME(X_l \mid X_k) = \bigcup_{x_{kj(k)} \in FC(X_k)} E(X_l \mid X_k = x_{kj(k)}) \]

“Laws” of Software Engineering

Principle of modularity (Parnas)

\[ M_p \] are module-induced clumps of decisions

\[ \forall p, i, k : X_i \in M_p, X_k \notin M_p, ME(X_i | X_k) = \emptyset \]

Conway’s Law

\[ T_c \] are team-induced clumps of decisions

\[ \forall p \exists c : M_p \subset T_c, \]

Additional Assumptions

- Constraint violation is binary
  - Decisions are either consistent or inconsistent
  - Satisfaction of functional requirements is binary
- Interdependencies are less troublesome when
  - Fewer people are involved in related decisions
  - People making related decisions communicate effectively
- Constraints are highly visible

Empirical Theory of Coordination

Technical Coordination Modeled as CSP

• Software engineering work = making decisions
• Constraint satisfaction problem
  • a project is a large set of mutually-constraining decisions, which are represented as
  • n variables $x_1, x_2, \ldots, x_n$ whose
  • values are taken from finite, discrete domains $D_1, D_2, \ldots, D_n$
  • constraints $p_k(x_{k1}, x_{k2}, \ldots, x_{kn})$ are predicates defined on
    the Cartesian product $D_{k1} \times D_{k2} \times \ldots \times D_{kj}$.
• Solving CSP is equivalent to finding an assignment for all variables that satisfies all constraints

Distributed Constraint Satisfaction

- Each variable $x_j$ belongs to one agent $i$
- Represented by relation $\text{belongs}(x_j, i)$
- Agents only know about a subset of the constraints
- Represent this relation as $\text{known}(P_l, k)$, meaning agent $k$ knows about constraint $P_l$
- Agent behavior determines global algorithm
- For humans, global behavior emerges
Measuring Coordination Requirements (C_R)  
(Constraints that span people)

<table>
<thead>
<tr>
<th>Concept</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Assignments</td>
<td>Task Dependencies</td>
</tr>
<tr>
<td>(A)</td>
<td>(D)</td>
</tr>
</tbody>
</table>
| a_{11} ... a_{1k}  
a_{n1} ... a_{nk} | d_{11} ... d_{1k}  
d_{k1} ... d_{kk} | a_{11} ... a_{in}  
a_{k1} ... a_{kn} |
| Developer modified files | Files changed together | Transpose of developer modified files |

Who needs to coordinate with whom

---

Volatility in Coordination Requirements

Measuring Congruence

Diff \( (C_R, C_A) \) = card \{ \text{diff}_{ij} | cr_{ij} > 0 \& ca_{ij} > 0 \} \\
Congruence \( (C_R, C_A) \) = Diff \( (C_R, C_A) \) / |C_R| \\

- Team structure
- Geographic location
- Use of chat
- On-line discussion

## Predicting Resolution Time

Table 2: Results from OLS Regression of Effects on Task Performance (*p < 0.10, *p < 0.05, **p < 0.01).

<table>
<thead>
<tr>
<th></th>
<th>Model I</th>
<th>Model II</th>
<th>Model III</th>
<th>Model IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.987**</td>
<td>3.631**</td>
<td>1.572*</td>
<td>1.751*</td>
</tr>
<tr>
<td>Dependency</td>
<td>0.897*</td>
<td>0.653*</td>
<td>0.784*</td>
<td>0.712*</td>
</tr>
<tr>
<td>Priority</td>
<td>-0.741*</td>
<td>-0.681*</td>
<td>-0.702*</td>
<td>-0.712*</td>
</tr>
<tr>
<td>Re-assignment</td>
<td>0.423*</td>
<td>0.487*</td>
<td>0.304*</td>
<td>0.324*</td>
</tr>
<tr>
<td>Customer MR</td>
<td>-0.730</td>
<td>-0.821</td>
<td>-0.932</td>
<td>-0.903</td>
</tr>
<tr>
<td>Release</td>
<td>-0.154*</td>
<td>-0.137*</td>
<td>-0.109*</td>
<td>-0.098*</td>
</tr>
<tr>
<td>Change Size (log)</td>
<td>1.542*</td>
<td>1.591*</td>
<td>1.428*</td>
<td>1.692*</td>
</tr>
<tr>
<td>Team Load</td>
<td>0.307*</td>
<td>0.317*</td>
<td>0.356*</td>
<td>0.374*</td>
</tr>
<tr>
<td>Programming Experience</td>
<td>-0.062*</td>
<td>-0.162*</td>
<td>-0.117*</td>
<td>-0.103*</td>
</tr>
<tr>
<td>Tenure</td>
<td>-0.269*</td>
<td>-0.265*</td>
<td>-0.239*</td>
<td>-0.248*</td>
</tr>
<tr>
<td>Component Experience (log)</td>
<td>-0.143*</td>
<td>-0.143*</td>
<td>-0.195*</td>
<td>-0.213*</td>
</tr>
<tr>
<td>Structural Congruence</td>
<td></td>
<td>-0.526*</td>
<td>-0.483*</td>
<td></td>
</tr>
<tr>
<td>Geographical Congruence</td>
<td></td>
<td>-0.317*</td>
<td>-0.312*</td>
<td></td>
</tr>
<tr>
<td>MR Congruence</td>
<td></td>
<td>-0.189*</td>
<td>-0.129*</td>
<td></td>
</tr>
<tr>
<td>IRC Congruence</td>
<td></td>
<td>-0.196*</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Interaction: ReleaseX Structural Congruence</td>
<td>0.007</td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction: ReleaseX Geographical Congruence</td>
<td>-0.013</td>
<td>-0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction: Release X MR Congruence</td>
<td>-0.009+</td>
<td>-0.011+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction: Release X IRC Congruence</td>
<td>-0.017*</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>809</td>
<td>809</td>
<td>1983</td>
<td>1983</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.787</td>
<td>0.872</td>
<td>0.756</td>
<td>0.854</td>
</tr>
</tbody>
</table>

(*p < 0.05, **p < 0.01)
Effects of Congruence

• Time to complete a work item is reduced by each of the types of congruence
  • Team structure congruence
  • Geographic location congruence
  • Chat congruence
  • On-line discussion congruence
Average Level of Congruence for Top 18 Contributors

- IRC
- MR
- Structure
- Geography
Average Level of Congruence for the Other 94 Developers

- **IRC**
- **MR**
- **Structure**
- **Geography**

The chart illustrates the average level of congruence across four releases (Release 1 to Release 4) for the mentioned categories.
The Story So Far . . .

- Focus on decisions and constraints
- Organization is solving a DCSP
- Ways of measuring constraints that span people
- Have the predicted effects
Decision Network Characteristics

- Centralized
- Small World
- Core/Periphery
- Density
- Scale Free
- Modular

Random network
Coordination: Five Propositions

- P1: Artifact design progresses by making decisions.
- P2: Decisions are linked by constraints in a potentially large and complex bipartite network.
  - The “constraint network”
- P3: The need for coordination among individuals arises from constraint network properties and assignment of decisions to people.
- P4: Coordination among individuals is the result of coordination actions, moderated by coordination capacity.
- P5: Coordination problems arise when coordination is insufficient for coordination needs.
Current View

Decision network structure

Work assignments

Coordination requirements

Coordination actions

Coordination success

Coordination actions
What Did Theory Do for Us?

- Explained the effects of modularity
- Led to measures of the need to coordinate
- Let us go beyond modularity and consider many different network structures and their impact
Barriers to Theory-based empirical research in SE

- Theory seen as mere decoration and distraction on top of statistical model
- Measures and constructs, not just variables
- Necessity to argue for practical application of each result
  - Dear Drs. Watson and Crick, I regret to inform you . . .
Collecting Additional Data

- Support exploratory analysis
- In what domains do explanations lie?
  - Social networks?
  - Information flow?
  - Processes?
- What are the important context variables?
- We should not passively accept whatever data happens to be available for other purposes . . .
- Hackystat
  - http://csdl.ics.hawaii.edu/Plone/research/hackystat
Connecting to Other Fields

- HCI, CSCW, Information Systems, Organizational Behavior, Management Science
Predictors of Conflict

+ 1. Revisions (article talk)
+ 2. Minor edits (article talk)
- 3. Unique editors (article talk)
+ 4. Revisions (article)
- 5. Unique editors (article)
+ 6. Anonymous edits (article talk)
- 7. Anonymous edits (article)

Regression model $R^2 \sim .9$
Wikipedia: Cost of Conflict and Coordination Growing

Quality and Contribution

- Many Wikipedia articles have quality ratings
- Quality as a function of
  - Number of editors
  - Concentration of editing activity
  - Communication
- Number of editors improves quality only if work is highly concentrated
- Communication improves quality when small number of editors, otherwise little effect
Connecting to Larger Community

- Will force us to look for general principles
- Better ways to test generality of results
- Ideas and techniques from other disciplines