Carnegie Mellon


# 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


Photo: I, MikeGogulski

# 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

Carnegie Mellon

# 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

Carnegie Mellon

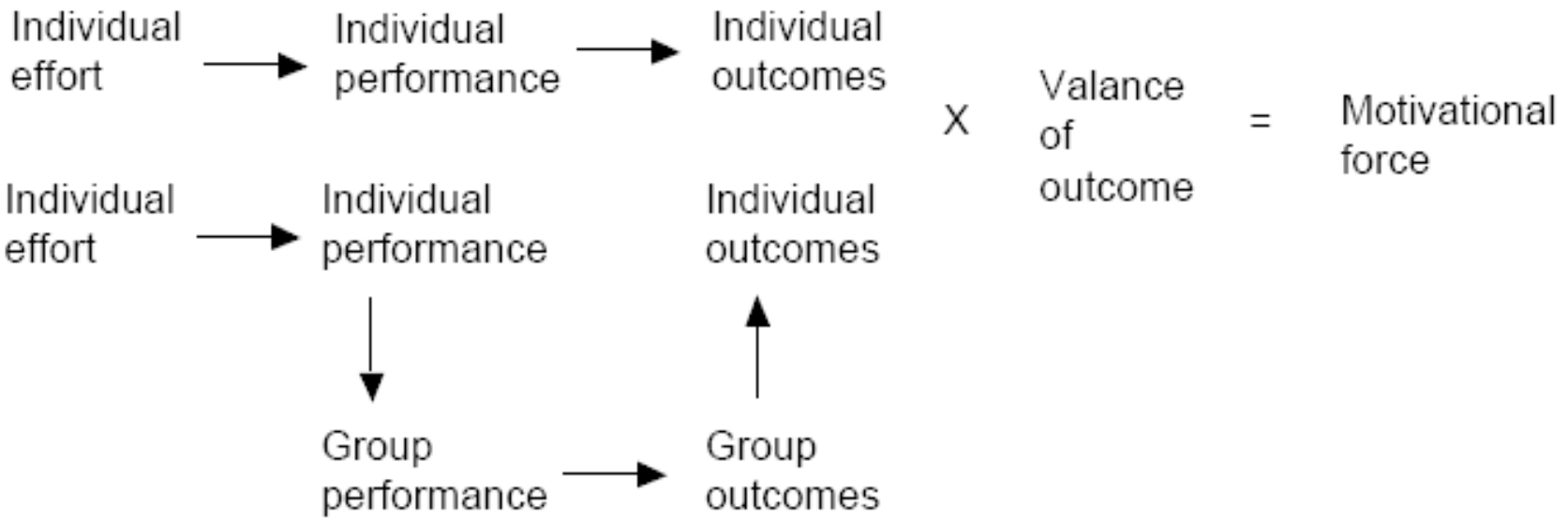
# A Definitive Review of Relevant Scientific Theories

# 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

Carnegie Mellon

# Social Psychology Theory:
# Collective Effort Model

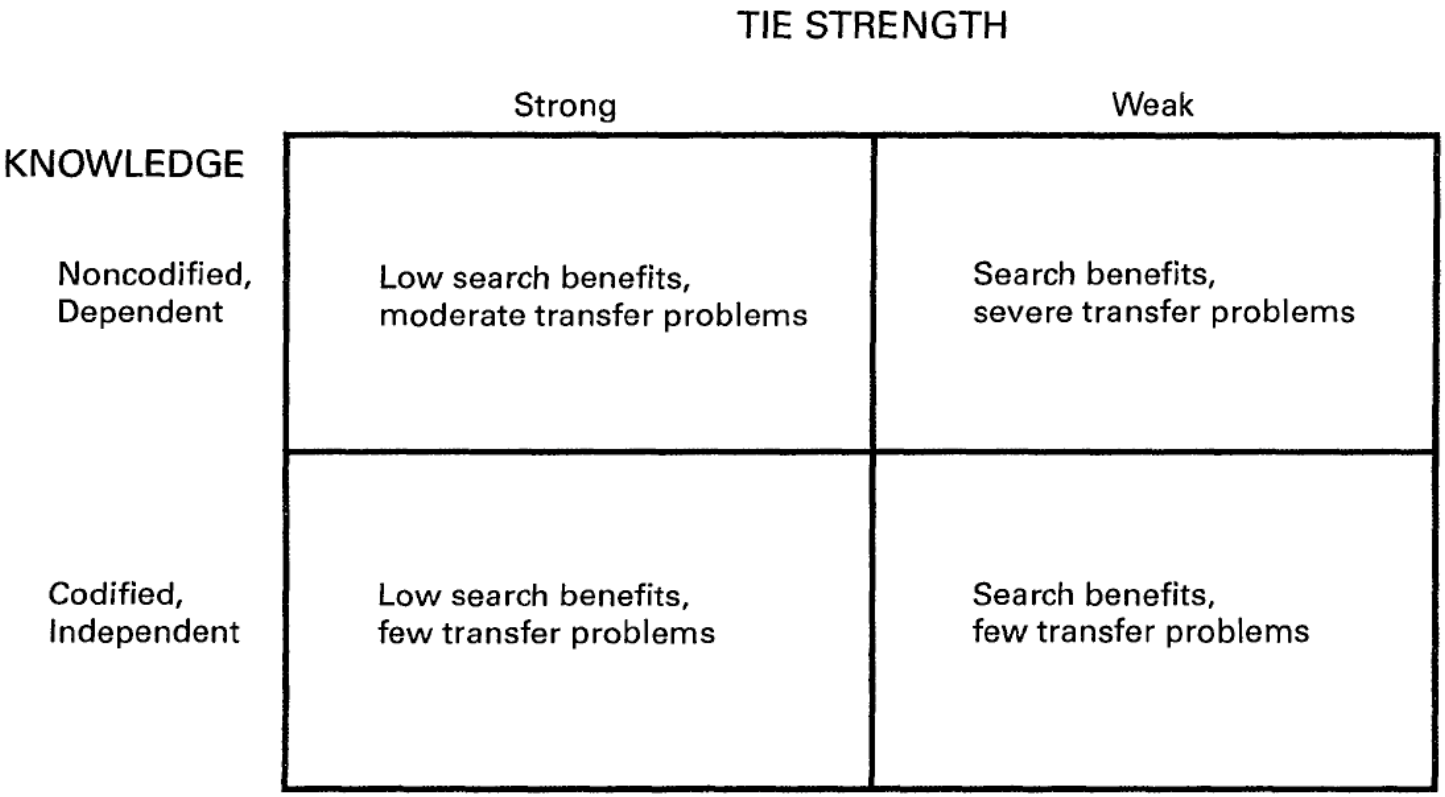





From Karau and Williams (2001) Understanding Individual Motivation in Groups: The Collective Effort Model. In Turner, M.E. (ed.), *Groups at Work: Theory and Research*. pp. 113-142

Carnegie Mellon

# Social Network Theory:
# Knowledge Transfer






Hansen, M.T. The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. *Administrative Science Quarterly,* Vol. 44, No. 1 (Mar., 1999), pp. 82-111

isr institute for SOFTWARE RESEARCH

# Theorizing about Coordination

- Collaborators
  - Beki Grinter
  - Audris Mockus
  - Marcelo Cataldo
  - Patrick Wagstrom
  - Kathleen Carley
  - Laura Dabbish
  - Anita Sarma

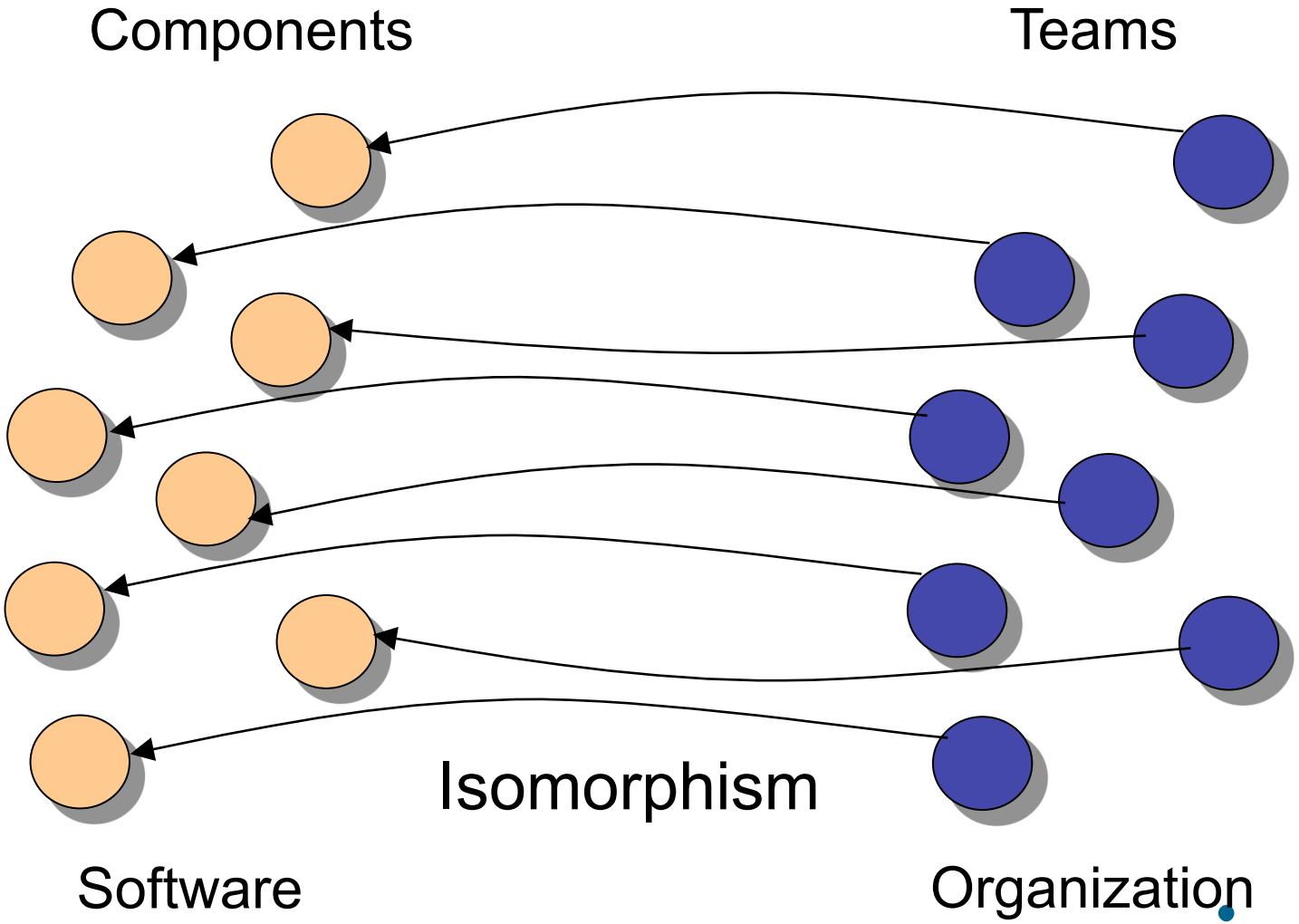
# Conway's Law

- "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**

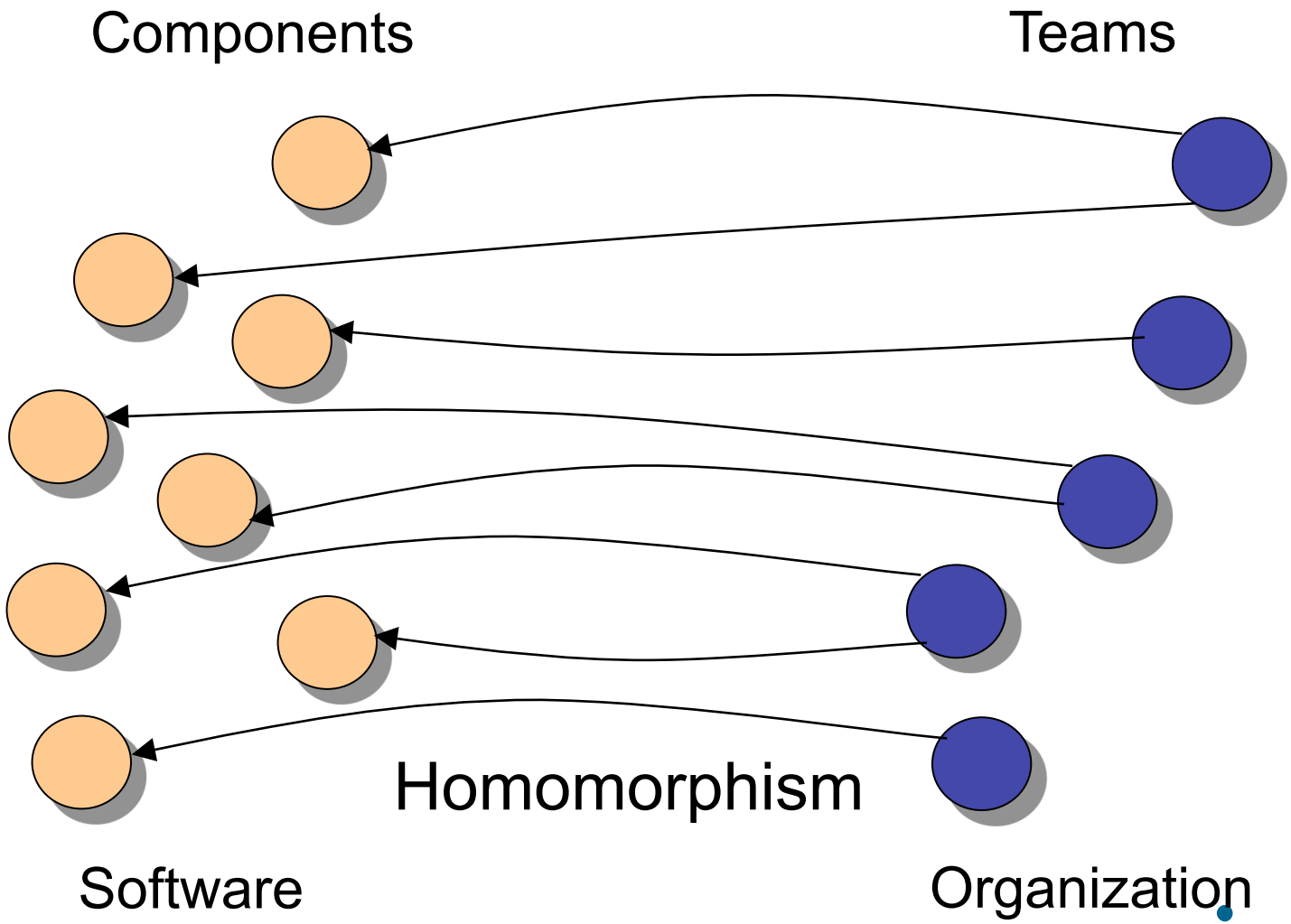


*M.E. Conway, "How Do Committees Invent?" Datamation, Vol. 14, No. 4, Apr. 1968, pp. 28–31.
**Baldwin, C. Y. and K. B. Clark (2000). *Design Rules: The Power of Modularity*. Cambridge, MA, The MIT Press.

# Conway's Law



# Conway's Law



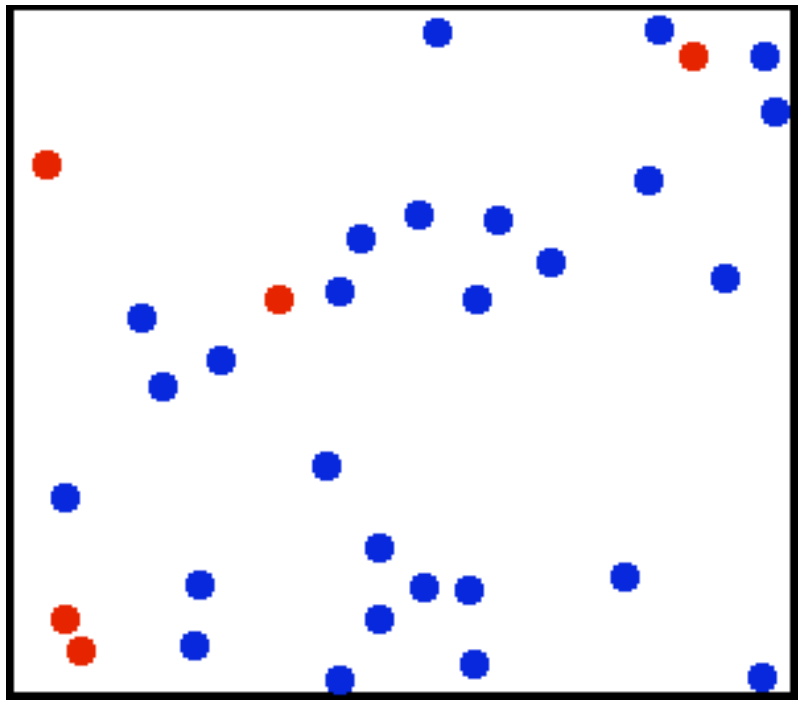
# Modularity: Just a Good Start

- 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?

Carnegie Mellon

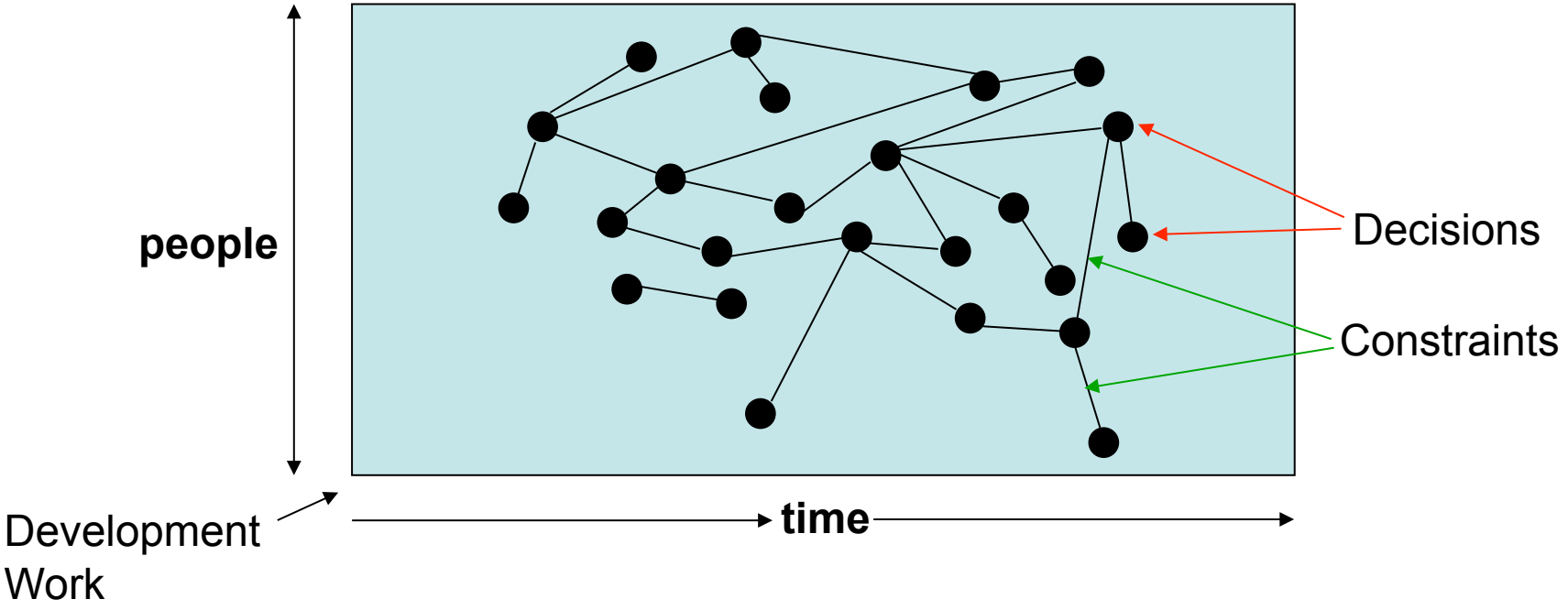
# Coordination and the Kinetic Theory of Gases





institute for SOFTWARE RESEARCH

# Software Development



Carnegie Mellon

# Key Definitions - 1

Project is a set of engineering decisions $X_i$

Feasibility function:

$$f(x_{1j(1)},\ldots,x_{nj(n)}) = \{1 \text{ iff product satisfies requirements,}$$
$$0 \text{ otherwise}\}$$

Feasible choices, $FC(X_i)$, is the set

$x_{ij*} : \forall k \neq i, \exists j(k)$ such that

$$f(x_{1j(1)},\ldots,x_{ij*},\ldots,x_{nj(n)}) = 1$$



Herbsleb, J.D. & Mockus, A. (2003). Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering. In proceedings, *ACM Symposium on the Foundations of Software Engineering (FSE), Helsinki, Finland, pp. 112-121.*

isr institute for SOFTWARE RESEARCH

# 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:

$$ME(X_l \mid X_k) = \cup_{x_{kj(k)} \in FC(X_k)} E(X_l \mid X_k = x_{kj(k)})$$



Herbsleb, J.D. & Mockus, A. (2003). Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering. In proceedings, *ACM Symposium on the Foundations of Software Engineering (FSE), Helsinki, Finland, pp. 112-121.*

Carnegie Mellon

# "Laws" of Software Engineering

Principle of modularity (Parnas)

$M_p$ are module-induced clumps of decisions

$$\forall p, i, k : X_i \in M_p, X_k \notin M_p, ME(X_i \mid X_k) = \emptyset$$

Conway's Law

$T_c$ are team-induced clumps of decisions

$$\forall p \exists c : M_p \subset T_c,$$



Herbsleb, J.D. & Mockus, A. (2003). Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering. In proceedings, *ACM Symposium on the Foundations of Software Engineering (FSE), Helsinki, Finland, pp. 112-121.*

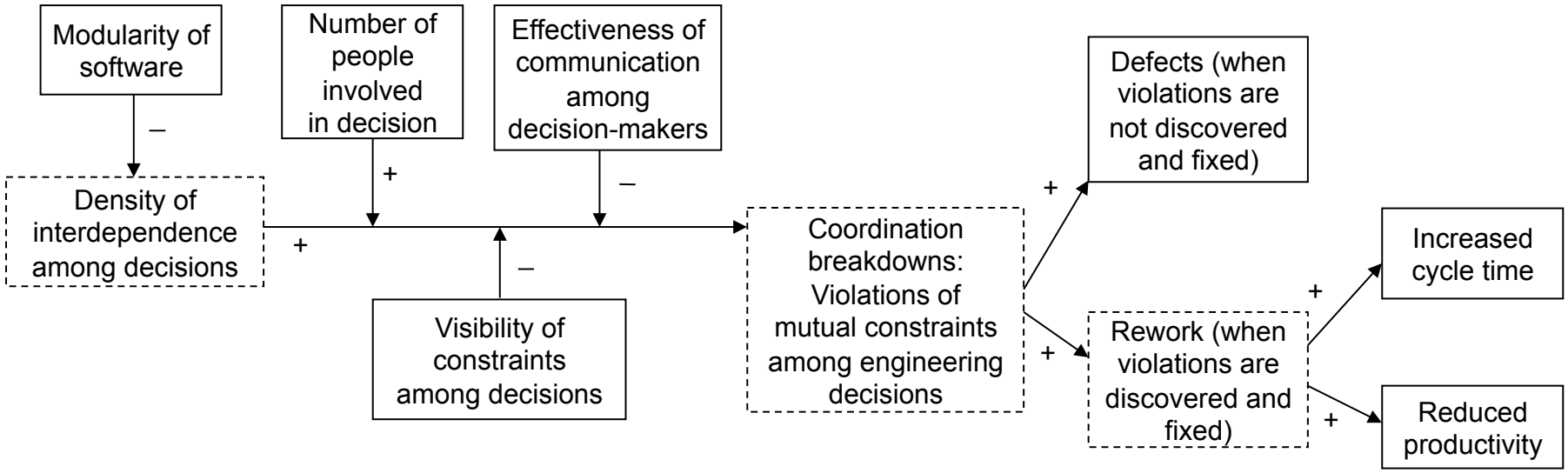
# Additional Assumptions

- 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



Herbsleb, J.D. & Mockus, A. (2003). Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering. In proceedings, *ACM Symposium on the Foundations of Software Engineering (FSE), Helsinki, Finland, pp. 112-121.*

# Empirical Theory of Coordination




Herbsleb, J.D. & Mockus, A. (2003). Formulation and Preliminary Test of an Empirical Theory of Coordination in Software Engineering. In proceedings, *ACM Symposium on the Foundations of Software Engineering (FSE), Helsinki, Finland, pp. 112-121.*

Carnegie Mellon

# 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 x1, x2, . . . , xn whose
  - values are taken from finite, discrete domains D1, D2, . . . , Dn
  - constraints pk(xk1, xk2, . . . , xkn) are predicates defined on
  - the Cartesian product Dk1 x DK2 x . . . x Dkj.
- Solving CSP is equivalent to finding an assignment for all variables that satisfies all constraints

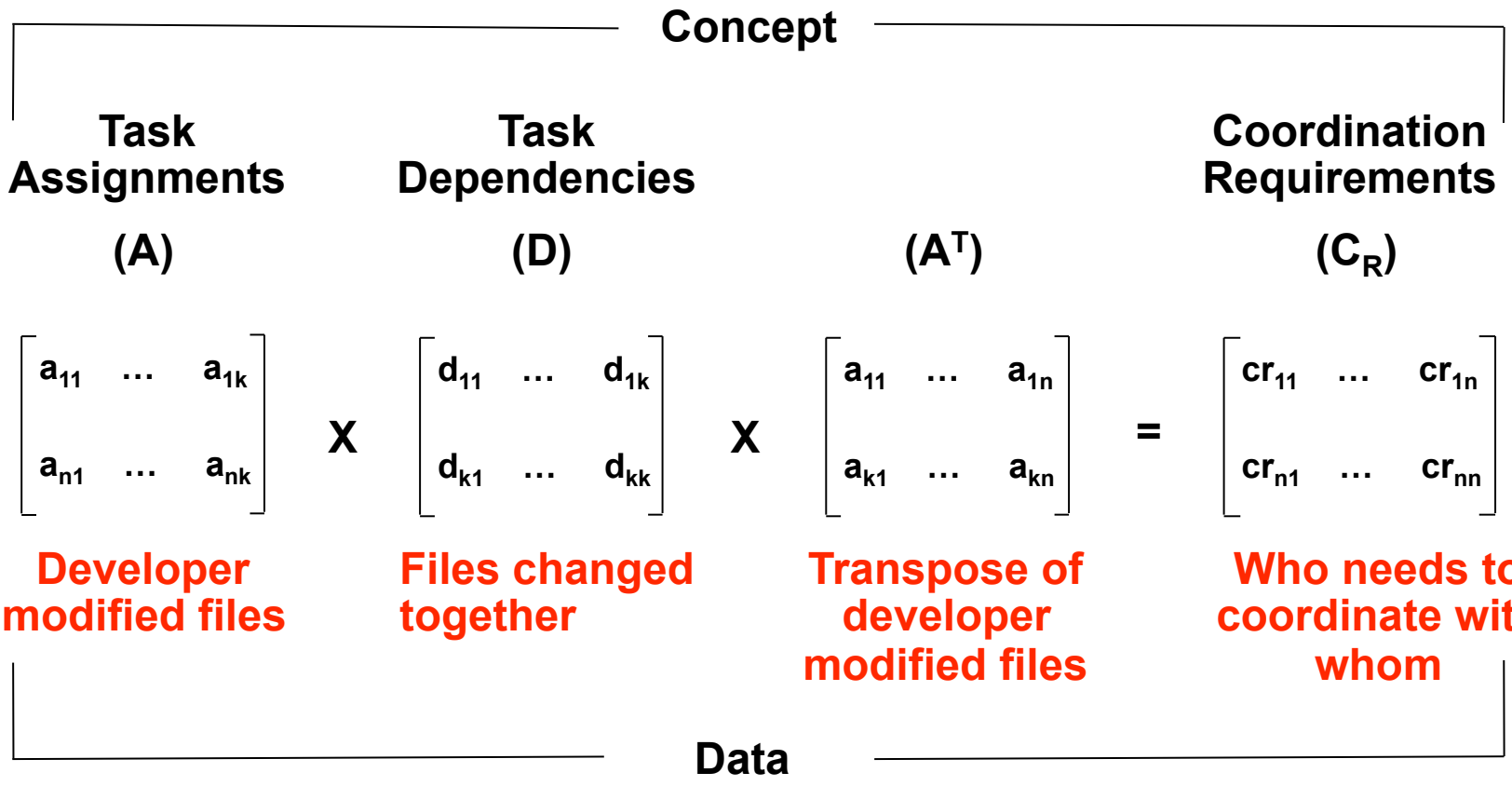
Formulation of CSP taken from Yokoo and Ishida, Search Algorithms for Agents, in G. Weiss (Ed.) *Multiagent Systems*, Cambridge, MA: MIT Press, 1999.

# Distributed Constraint Satisfaction

- Each variable xj belongs to one agent i
- Represented by relation belongs(xj,i)
- Agents only know about a subset of the constraints
- Represent this relation as known(Pl, k), meaning agent k knows about constraint Pl
- Agent behavior determines global algorithm
- For humans, global behavior emerges

Carnegie Mellon

# Measuring Coordination Requirements ($C_R$) (Constraints that span people)



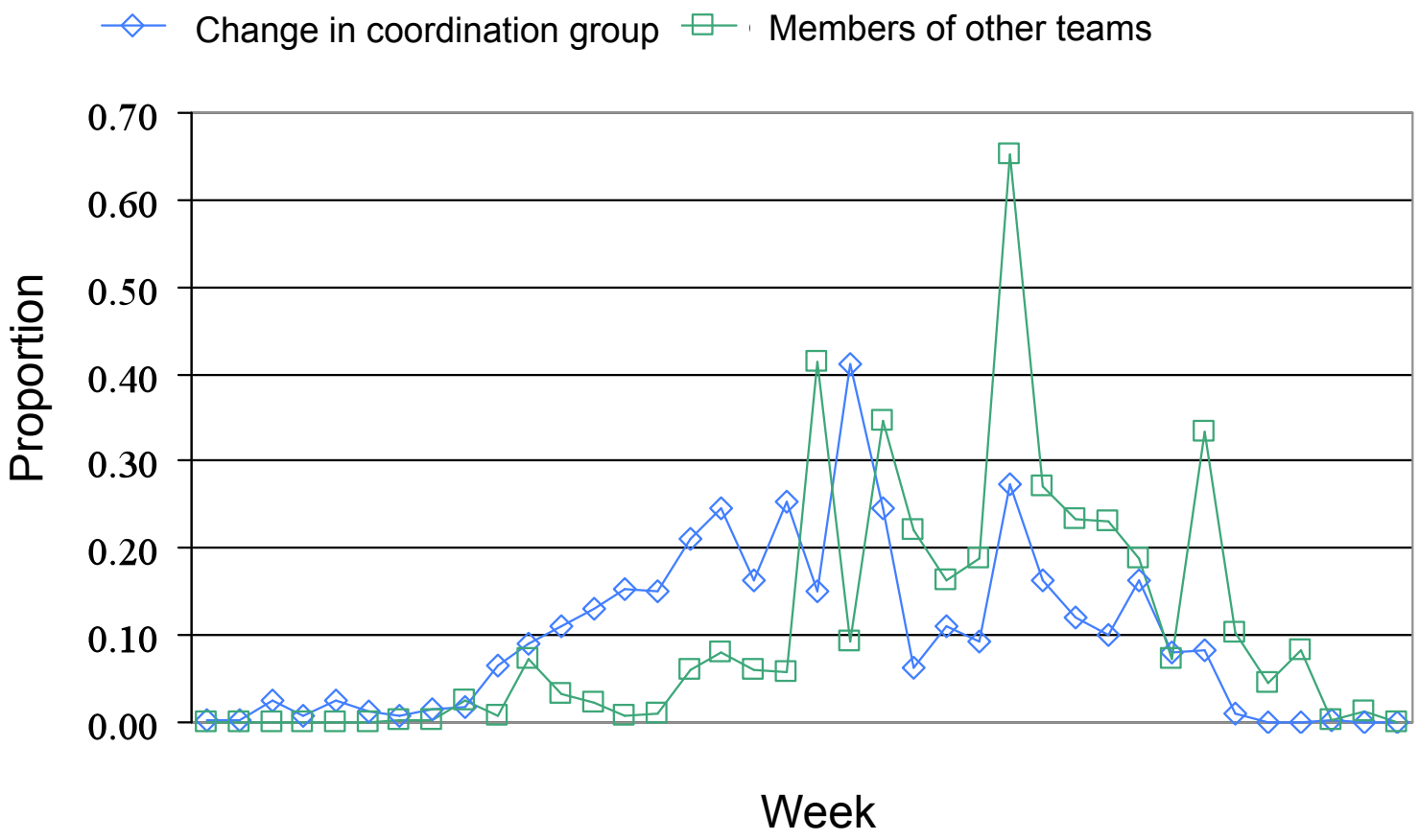

Cataldo, M., Wagstrom, P., Herbsleb, J.D., Carley, K. (2006). Identification of coordination requirements: Implications for the design of collaboration and awareness tools. In Proceedings, *ACM Conference on Computer-Supported Cooperative Work, Banff Canada,* pp. 353-362.


29

institute for SOFTWARE RESEARCH

Carnegie Mellon

# Volatility in Coordination Requirements




Cataldo, M., Wagstrom, P., Herbsleb, J.D., Carley, K. (2006). Identification of coordination requirements: Implications for the design of collaboration and awareness tools. In Proceedings, *ACM Conference on Computer-Supported Cooperative Work, Banff Canada,* pp. 353-362.



institute for SOFTWARE RESEARCH

# Measuring Congruence

**Coordination Requirements ($C_R$)**

$$\begin{bmatrix} cr_{11} & \dots \\ cr_{1n} & \\ cr_{n1} & \dots \\ cr_{nn} & \end{bmatrix}$$

$\Longleftrightarrow$

**Actual Coordination ($C_A$)**

$$\begin{bmatrix} ca_{11} & \dots \\ ca_{1n} & \\ ca_{n1} & \dots \\ ca_{nn} & \end{bmatrix}$$

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

$$\textit{Diff}\ (C_R, C_A) = \textit{card}\ \{\ \textit{diff}_{ij} \mid \textit{cr}_{ij} > 0\ \&\ \textit{ca}_{ij} > 0\ \}$$

$$\textit{Congruence}\ (C_R, C_A) = \textit{Diff}\ (C_R, C_A) / |C_R|$$

Cataldo, M., Wagstrom, P., Herbsleb, J.D., Carley, K. (2006). Identification of coordination requirements: Implications for the design of collaboration and awareness tools. In Proceedings, *ACM Conference on Computer-Supported Cooperative Work, Banff Canada,* pp. 353-362.


31

Carnegie Mellon

# Predicting Resolution Time

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

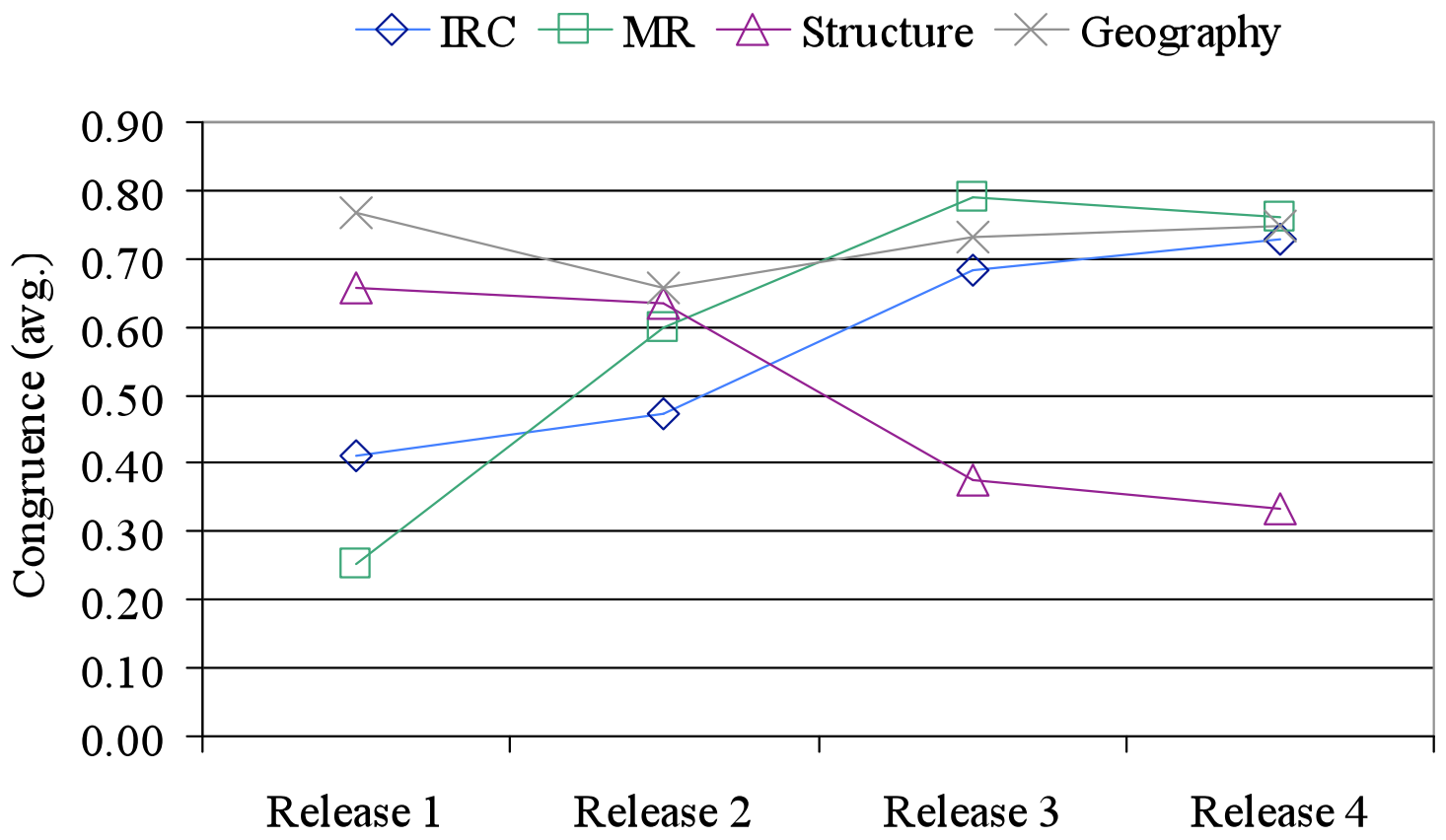
| | Model I | Model II | Model III | Model IV |
|---|---|---|---|---|
| *(Intercept)* | 2.987** | 3.631** | 1.572* | 1.751* |
| *Dependency* | 0.897* | 0.653* | 0.784* | 0.712* |
| *Priority* | -0.741* | -0.681* | -0.702* | -0.712* |
| *Re-assignment* | 0.423* | 0.487* | 0.304* | 0.324* |
| *Customer MR* | -0.730 | -0.821 | -0.932 | -0.903 |
| *Release* | -0.154* | -0.137* | -0.109* | -0.098* |
| *Change Size (log)* | 1.542* | 1.591* | 1.428* | 1.692* |
| *Team Load* | 0.307* | 0.317* | 0.356* | 0.374* |
| *Programming Experience* | -0.062* | -0.162* | -0.117* | -0.103* |
| *Tenure* | -0.269* | -0.265* | -0.239* | -0.248* |
| *Component Experience (log)* | -0.143* | -0.143* | -0.195* | -0.213* |
| ***Structural Congruence*** | | **-0.526*** | | **-0.483*** |
| ***Geographical Congruence*** | | **-0.317*** | | **-0.312*** |
| ***MR Congruence*** | | **-0.189*** | | **-0.129*** |
| ***IRC Congruence*** | | **-0.196*** | | **--** |
| ***Interaction: ReleaseX Structural Congruence*** | | **0.007** | | **0.009** |
| ***Interaction:ReleaseXGeographical Congruence*** | | **-0.013** | | **-0.017** |
| ***Interaction: Release X MR Congruence*** | | **-0.009$^{+}$** | | **-0.011$^{+}$** |
| ***Interaction: Release X IRC Congruence*** | | **-0.017*** | | **--** |
| N | 809 | 809 | 1983 | 1983 |
| Adjusted $R^2$ | 0.787 | 0.872 | 0.756 | 0.854 |

(* p < 0.05, ** p < 0.01)

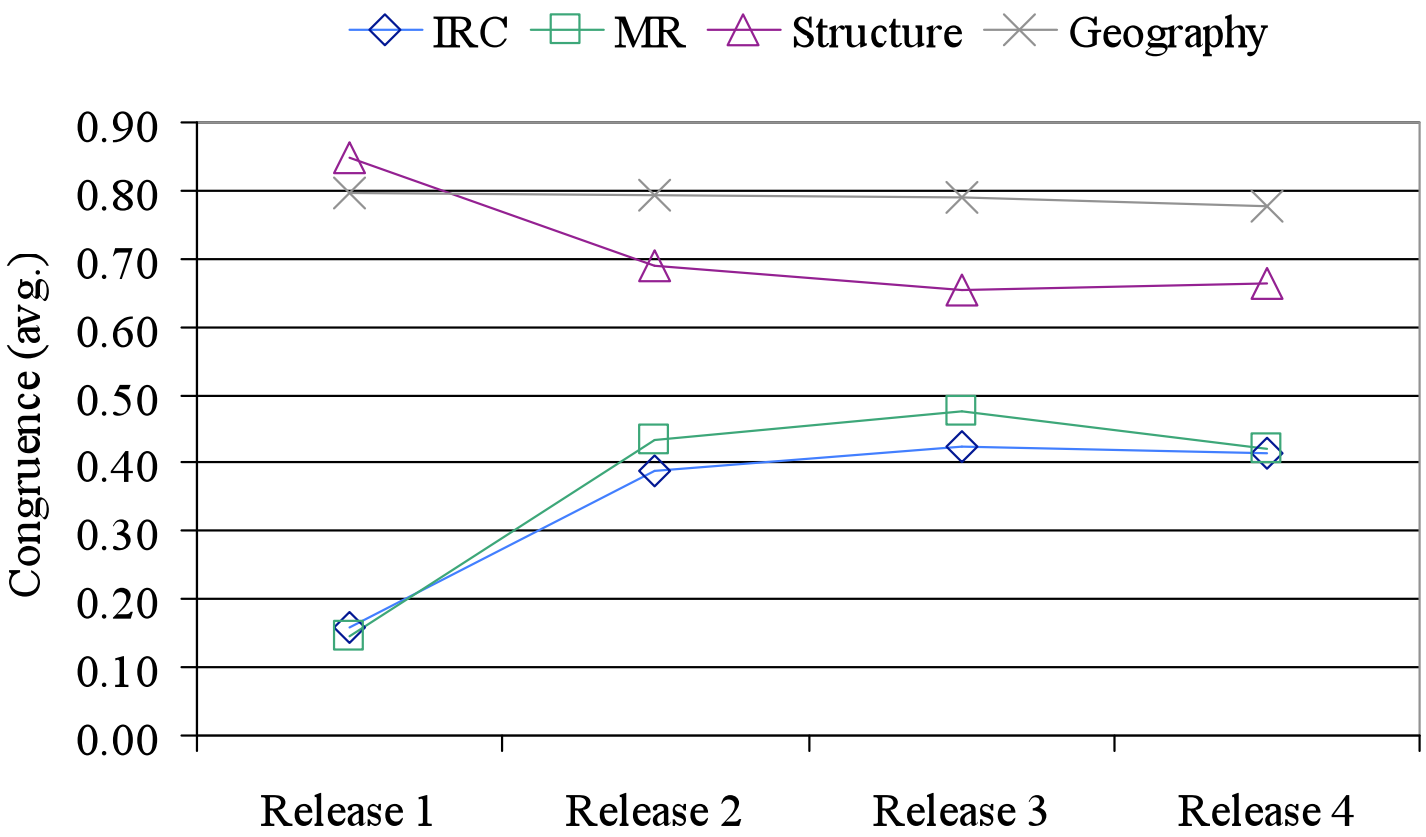
# Effects of Congruence

- 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



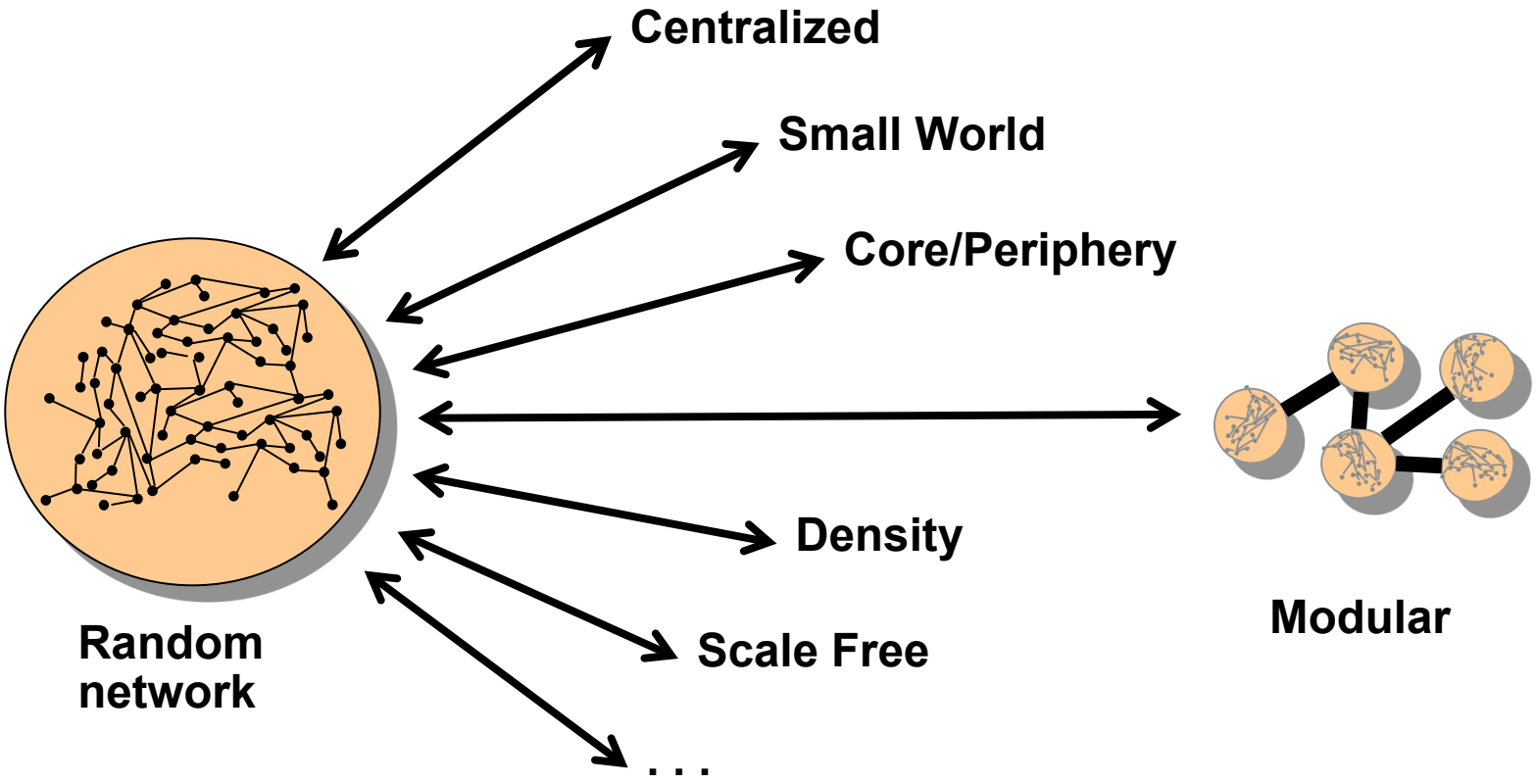
# Average Level of Congruence for the Other 94 Developers



# 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

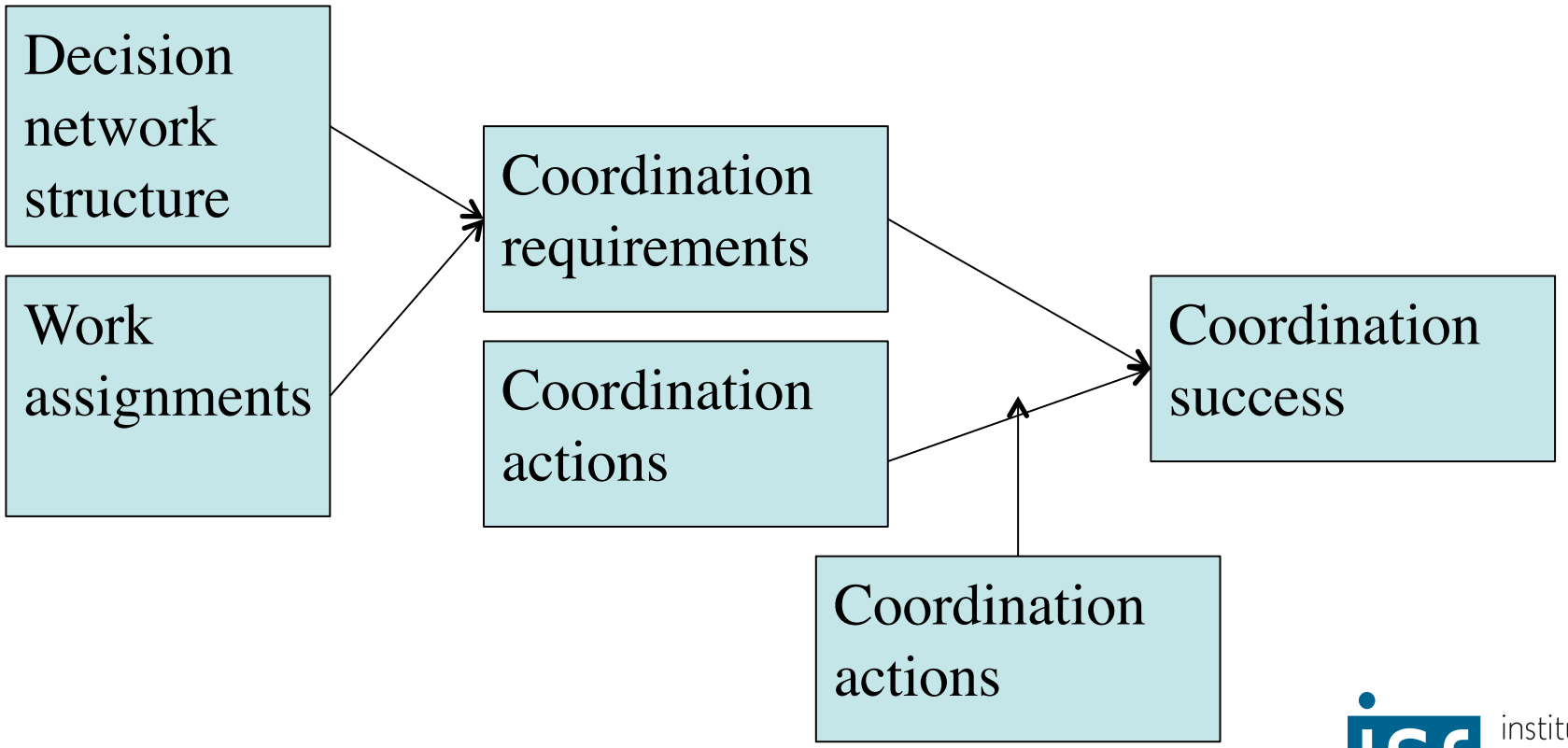
# Decision Network Characteristics



# 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



# 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
- Example topic: Wikipedia

Carnegie Mellon

# 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)

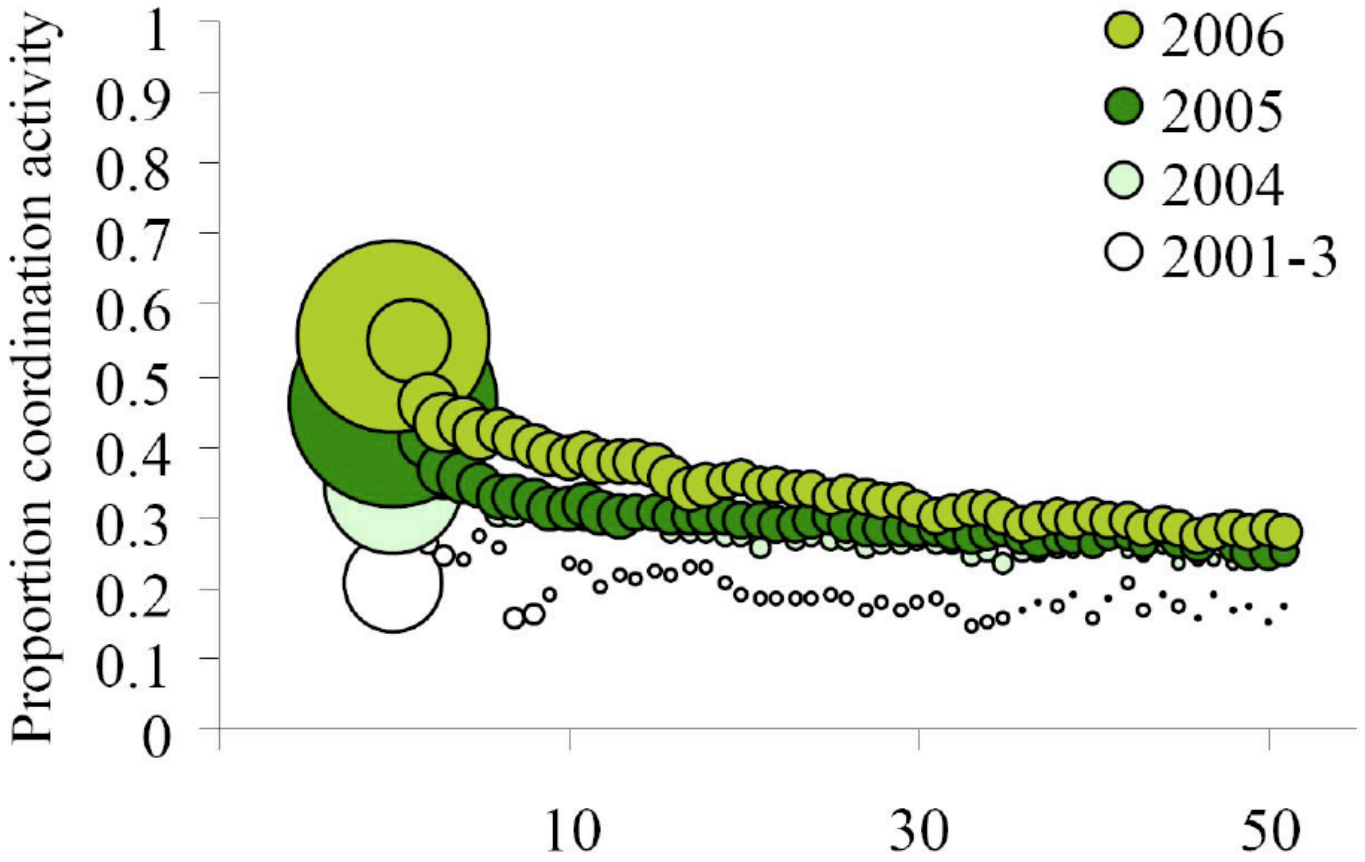
Regression model $R^2$ ~.9



Kittur, et al, He Says, She Says: Conflict and Coordination in Wikipedia, *CHI* 2007

Carnegie Mellon

# Wikipedia: Cost of Conflict and Coordination Growing






Kittur, A. & Kraut, R.E. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. CSCW 2008.

institute for SOFTWARE RESEARCH

# 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