Using Relational Topic Models to Capture Coupling among Classes in Object-Oriented Software Systems

Malcom Gethers and Denys Poshyvanyk

26th IEEE International Conference on Software Maintenance
Timişoara, Romania
September 16, 2010

Coupling

- Coupling is a fundamental property of software design
- Coupling metrics quantify the degree of relationship between software components
- Applications: impact analysis, defect prediction, software re-modularization

Prior Work

- Structural Coupling Metrics
  - Coupling between classes (CBO) [Chidamber'04]
  - Response for class (RFC) [Chidamber'04]
  - Message passing coupling (MPC) [Li'93]
  - Data abstraction coupling (DAC) [Li'93]
  - Information-flow based coupling (IPC) [Lee'95]
  - A suite of coupling measures by Briand et al: ACAIC, OCAIC, ACMIC and OCMIC
  - …
- Evolutionary Coupling Metrics
  - Logical Coupling [Gall'03][Zimmermann'05]
- Conceptual Coupling Metrics
  - Conceptual Coupling of Classes (CoCC) [Poshyvanyk'09]

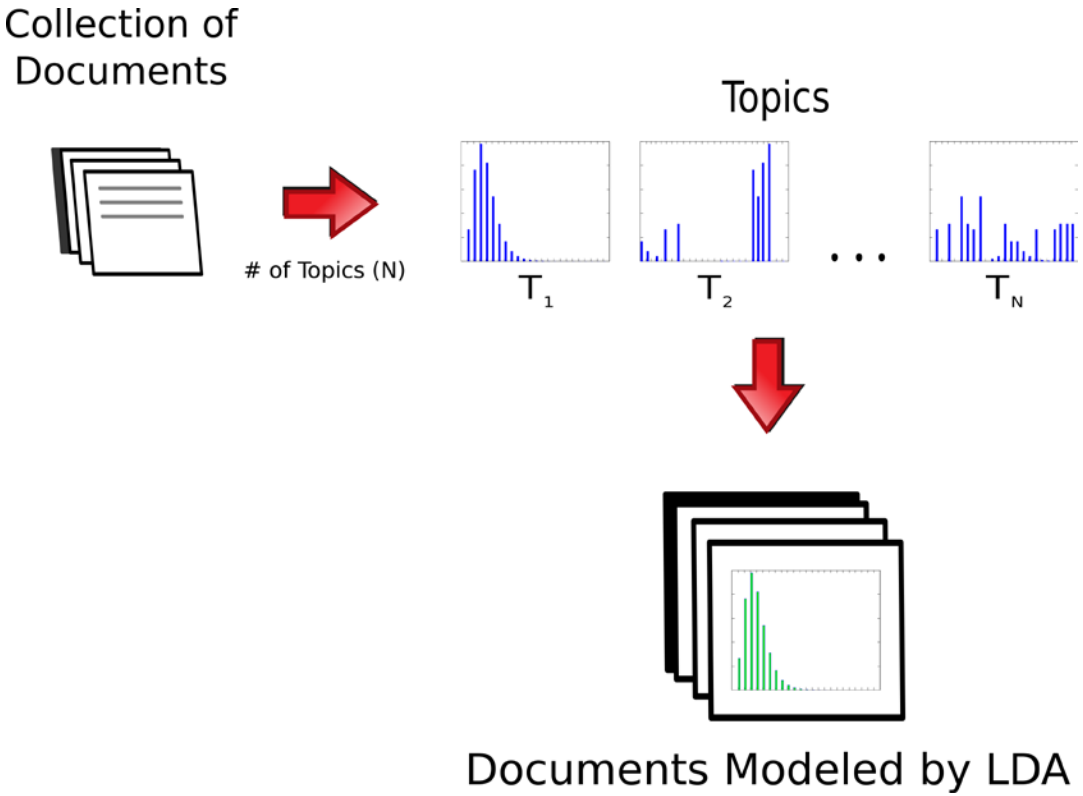
Problem

- Many existing coupling metrics are based on structural information with very few metrics which use textual information to capture coupling

Goal

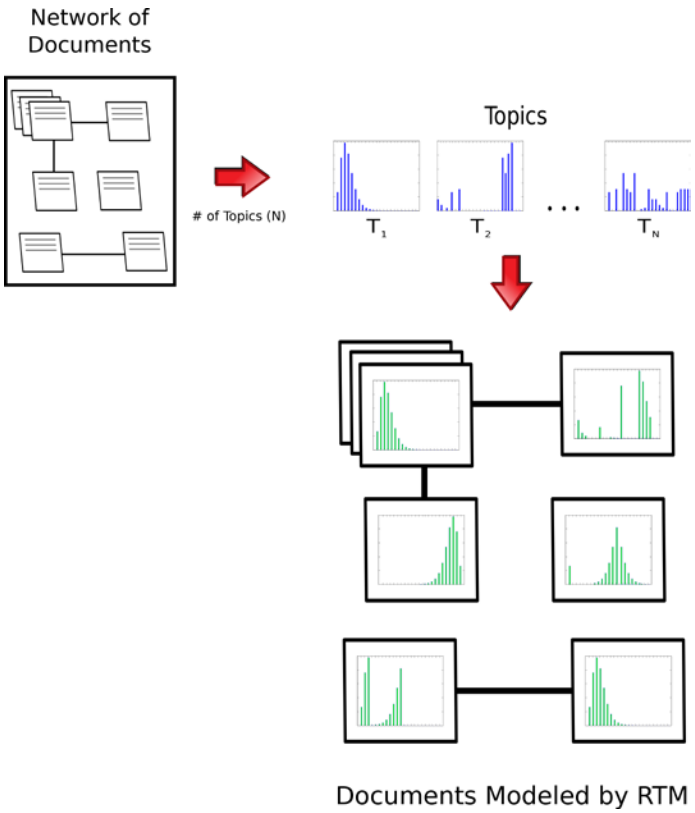
- Define a novel conceptual coupling metric based on advanced Information Retrieval (IR) techniques (i.e., Relational Topic Models)
- Show that the conceptual coupling metrics are useful for measuring the degree of interaction/relationship between classes in Object-Oriented Systems

Topic Models



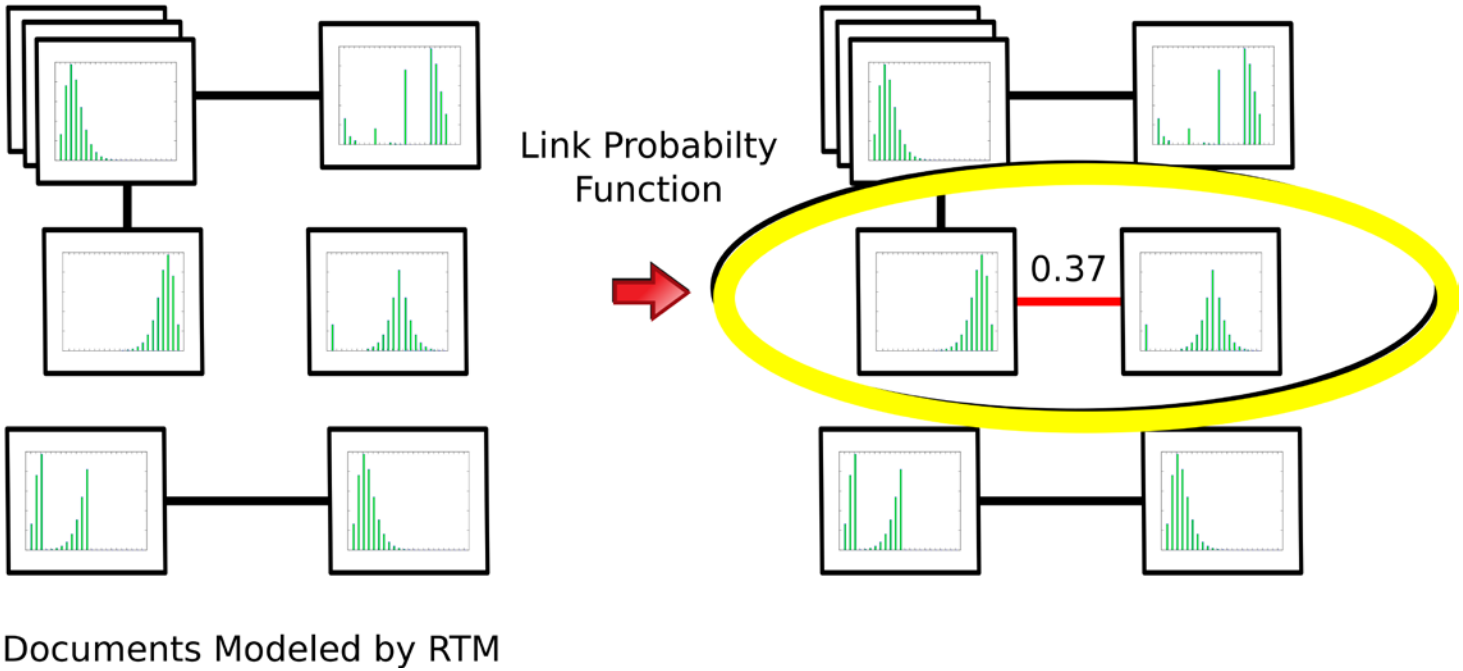
- Probabilistic Topic Models (Latent Dirichlet Allocation – LDA [Blei'03])
- Models documents as mixtures of topics

Relational Topic Model (RTM)



- Relational Topic Model [Chang'10] - Advance topic model
- Models existing and predicts new relationships between documents

Relational Topic Model (RTM)

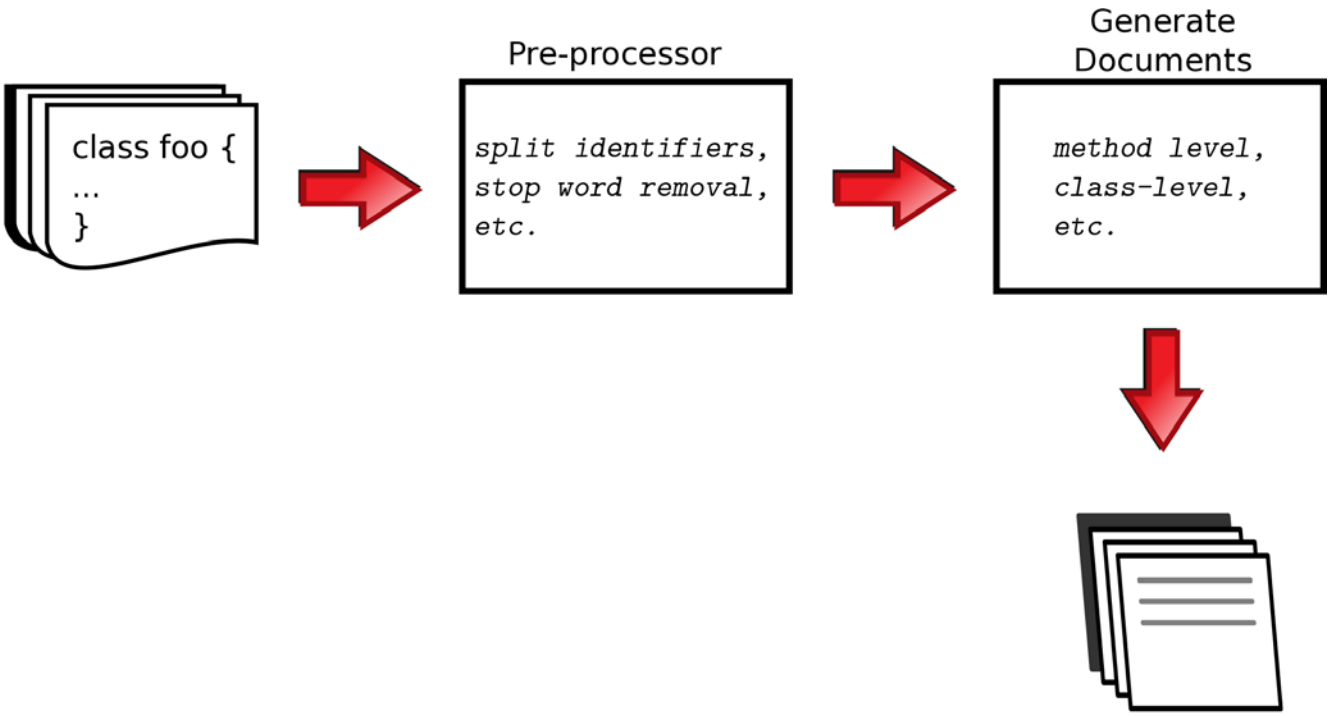


- Link Probability Function – utilizes document text and known links
- Applications: suggest citations, identify friends, predict web pages [Chang'10]

Benefits of RTM

- Topic Model - models documents as probabilistic mixtures of topics
- Models relationships between documents
- Link Probability Function - indicates how likely it is that a link exists between two documents
- Flexibility - capable of making predictions with and without known links

Applying Topic Models to Source Code



- Generate a text corpus of documents when provided a software system as an input

Measuring Coupling using RTM

Relational Topic Based Coupling between Classes:

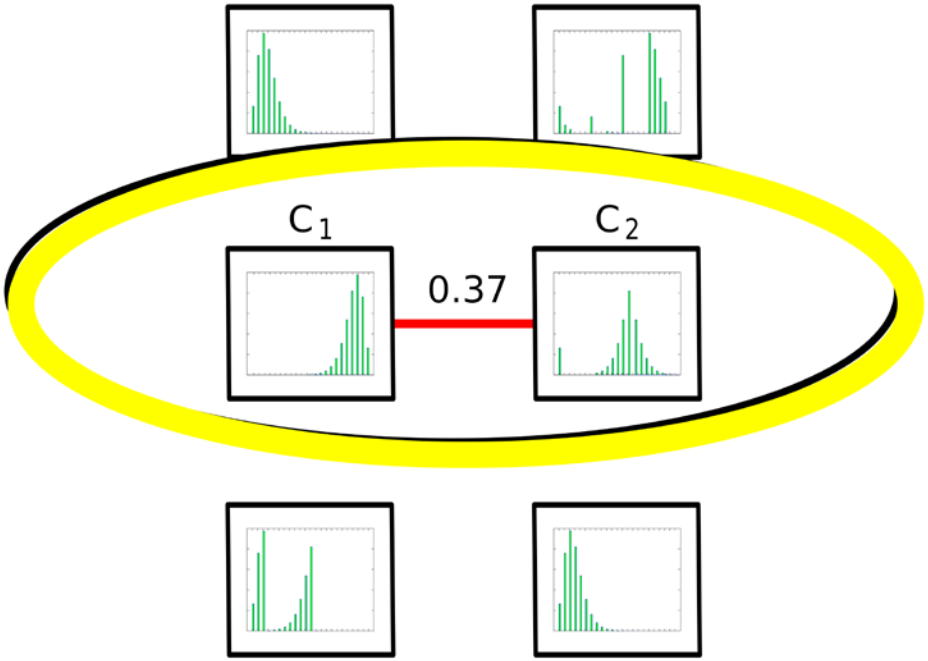
$$RTC_{CLASS}(C_1, C_2) = RTM(C_1, C_2)$$

Relational Topic Based Coupling (system-level):

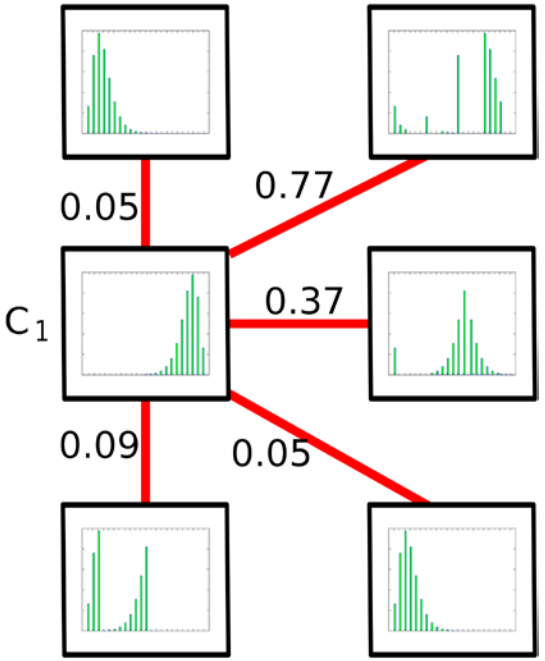
$$RTC_{SYSTEM}(C_i) = \frac{\sum_{j \in C}^{n} RTC(C_i, C_j)}{n}$$

Measuring Coupling using RTM

$RTC_{CLASS}(C_1,C_2) = 0.37$



$RTC_{SYSTEM}(C_1) = 0.27$



- Predicts links between classes in a given software system

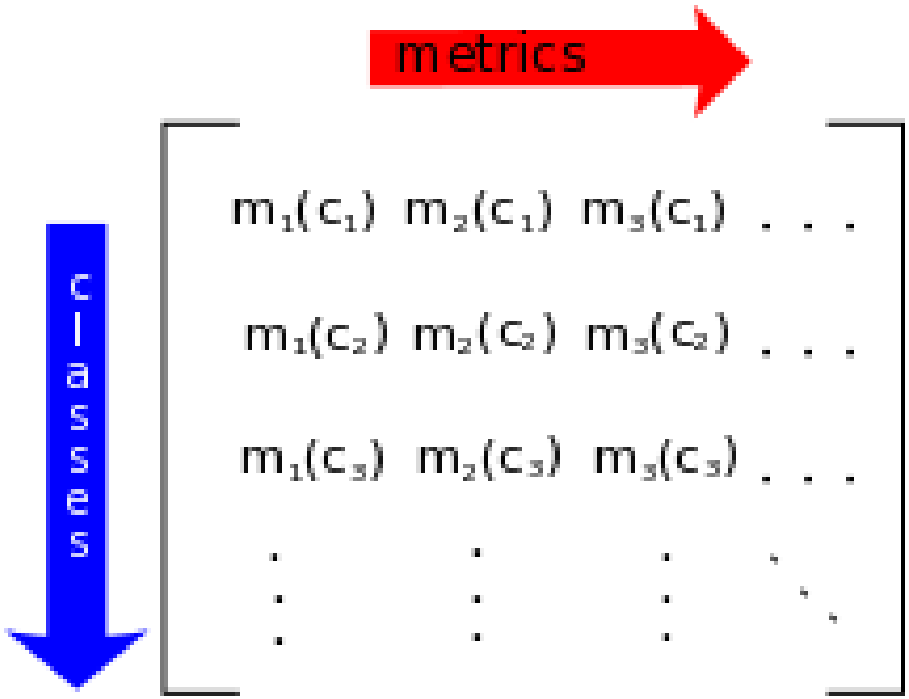
Case Study

Is RTC a useful and meaningful coupling metric?

Settings of Case Study

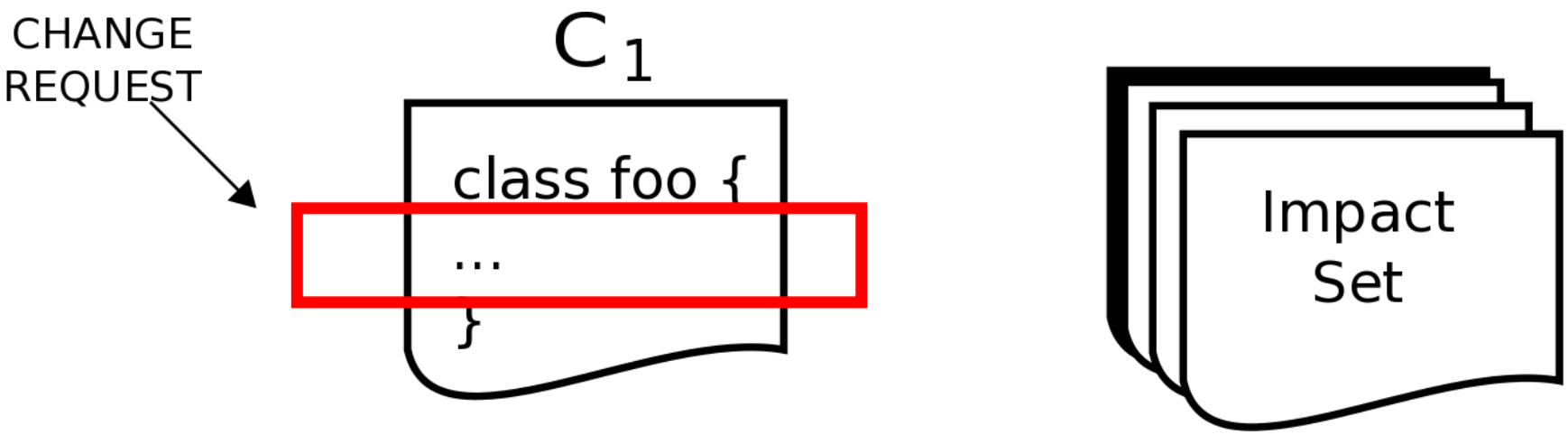
- Subject systems: 11 C/C++ and two Java systems
- Metrics: CBO, RFC, MPC, DAC, ICP, ACAIC, CAIC, ACMIC, OCMIC, and CoCC.
- Tools: Columbus [Ferenc'04] and IRC^{2}M [Posyvanyk'06]
- Initial analysis: Principal Component Analysis (PCA)
- Task: Impact Analysis (IA) [Briand'99]

Case Study



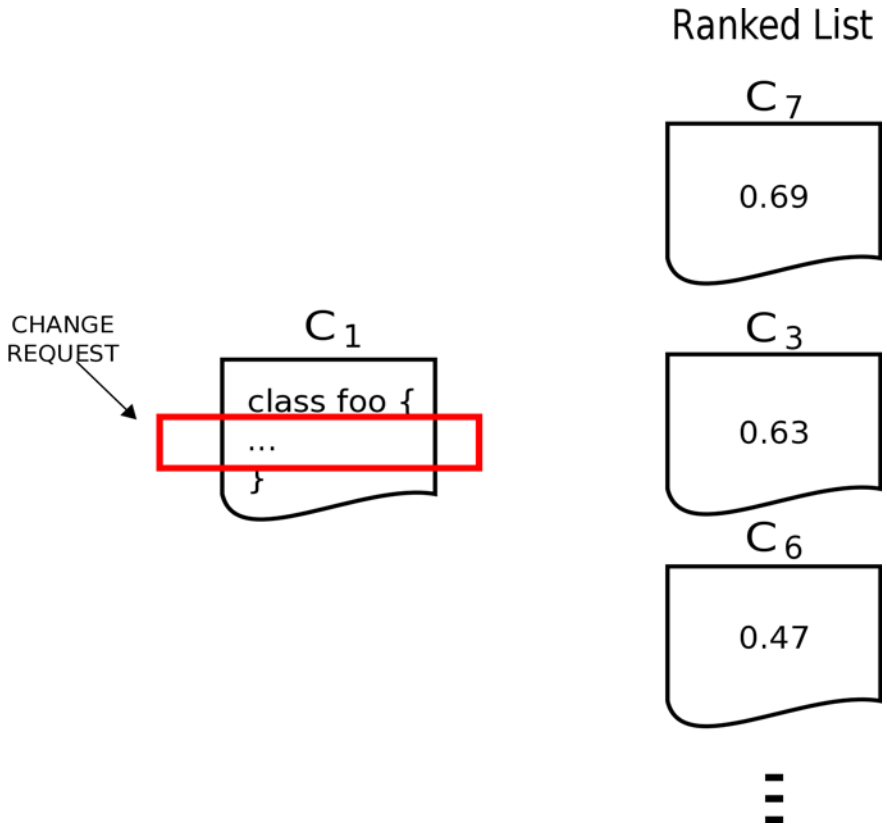
- Principal Component Analysis of coupling metrics

Case Study



- Impact Analysis

Case Study



- Impact Analysis using Coupling Metrics [Briand'99]
- Ground Truth - Bug reports/Revision History
- Precision/Recall

Case Study – RQ1

RQ1: Is RTC_{SYSTEM} metric distinct when compared to existing structural and conceptual coupling metrics?

- Principal Component Analysis
- Software Systems - 11 C/C++ software systems [Poshyvanyk'09]

Case Study – RQ1

	PC1	PC2	PC3	PC4	PC5	PC6
Proportion	29.83%	16.65%	11.76%	16.44%	8.17%	8.11%
Cumulative	29.83%	46.49%	58.25%	74.69%	82.85%	90.97%
RTC_{SYSTEM}	0.02	0.23	0.25	0.01	0.00	**0.93**
CoCC	-0.03	0.29	**0.85**	-0.05	0.23	0.15
$CoCC_{max}$	0.33	-0.23	**0.75**	0.07	-0.24	0.19
CBO	**0.83**	0.21	0.19	0.27	0.09	0.01
RFC	**0.88**	0.02	0.01	0.19	0.15	0.10
MPC	**0.95**	0.03	0.03	0.15	0.07	-0.02
ICP	**0.89**	0.13	0.11	0.20	0.13	-0.03
ACAIC	0.11	**0.91**	-0.03	0.17	0.09	0.16
ACMIC	0.12	**0.91**	0.12	0.11	0.08	0.09
DAC	0.31	0.22	0.00	**0.91**	0.12	0.02
OCAIC	0.29	0.09	0.01	**0.93**	0.13	0.00
OCMIC	0.32	0.15	0.04	0.23	**0.88**	0.00

- Principal Component Analysis Results

Case Study – RQ2

RQ2: Does RTC_{class} outperform existing structural metrics for the task of impact analysis?

- Impact Analysis
- Software System – Mozilla v1.6 [Posyvanyk'09]

Case Study – RQ2

	10		20		30		40		50		100		200		500	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
RTC_{CLASS}	17.3	14.0	15.8	22.0	14.7	27.3	13.8	31.6	13.1	36.1	9.5	47.4	6.8	59.1	4.1	75.9
$CCBC_{max}$	27.8	14.6	24.7	22.1	18.4	34.5	18.4	34.5	18.4	34.5	12.6	43.1	8.4	52.4	4.6	65.1
ICP	11.9	6.9	10.1	9.7	8.6	16.5	8.6	16.5	8.6	16.5	6.5	22.8	4.13	27.3	2.6	39.0
PIM	11.3	6.6	9.8	9.6	8.5	16.3	8.5	16.3	8.5	16.3	6.5	22.6	4.1	27.1	2.6	38.9
CCBC	10.8	5.6	9.5	8.9	6.7	14.1	6.7	14.1	6.7	14.1	5.2	19.8	4.0	27.0	3.0	44.8
CBO	7.2	6.2	5.4	9.4	2.8	11.3	2.8	11.3	2.8	11.3	1.6	12.0	1.05	13.2	1.0	26.5
MPC	6.6	5.7	3.9	6.7	1.7	7.0	1.7	7.0	1.7	7.0	0.9	7.2	0.7	8.6	1.2	22.7
OCMIC	2.0	2.1	1.1	2.2	0.5	2.3	0.5	2.3	0.5	2.3	0.3	2.5	0.5	4.3	1.2	20.2
OCAIC	1.7	2.0	1.0	2.1	0.4	2.1	0.4	2.1	0.4	2.1	0.2	2.3	0.5	4.2	1.2	19.9
DAC	1.8	2.0	1.0	2.1	0.4	2.1	0.4	2.1	0.4	2.1	0.2	2.3	0.2	2.4	0.2	2.4
ACMIC	0.9	0.4	0.5	0.4	0.2	0.4	0.2	0.4	0.2	0.4	0.1	0.6	0.4	2.4	1.2	18.5
ACAIC	0.8	0.3	0.4	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.1	0.5	0.4	2.4	1.2	18.5

- Impact Analysis: P=Precision R=Recall

Case Study – RQ3

RQ3: Does RTC_{class} or its combinations with conceptual coupling metrics outperform existing coupling metrics for the task of impact analysis?

- Impact Analysis
- Software System – Eclipse v3.0 and Rhino v1.5R6
- Baseline - $CCBC_{max}$
- Affine Transformation (equal weight to both techniques)

$$RTC_{CLASS} + CCBC_{\max}(C_i, C_j) =$$

$$\lambda \times norm(RTC_{CLASS}(C_i, C_j)) + (1-\lambda) \times norm(CCBC_{\max}(C_i, C_j))$$

Case Study – RQ3

		10		20		30		40		50		100		200		500	
		P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
Eclipse	$RTC_{CLASS} + CCBC_{max}$	21	25	17	35	15	40	13	45	12	49	9	61	5	70	3	76
	$CCBC_{max}$	19	23	15	31	13	36	12	41	11	44	8	56	5	67	3	72
	Absolute gain	2	2	2	4	2	4	1	4	1	5	1	5	0	4	0	4
	Relative gain	11	9	13	13	15	11	8	10	9	11	13	9	0	4	0	6
Rhino	$RTC_{CLASS} + CCBC_{max}$	14	38	11	57	9	68	8	75	7	79	98	2	n/a	n/a	n/a	n/a
	$CCBC_{max}$	13	32	11	52	9	64	8	73	6	77	98	2	n/a	n/a	n/a	n/a
	Absolute gain	1	6	0	5	0	4	0	2	1	2	0	0	n/a	n/a	n/a	n/a
	Relative gain	8	19	0	10	0	6	0	3	17	3	0	0	n/a	n/a	n/a	n/a

- Impact Analysis: P=Precision R=Recall
- Note: Wilcoxon test confirms improvement statistically significant for $p = 0.05$.

Threats to Validity

- Set of software systems
- Compared RTC to structural and conceptual metrics
- Conceptual metrics depend on coherent naming conventions
- Bug reports used to measure accuracy of impact analysis

Conclusion

- Defined a novel coupling metric based on Relational Topic Model
- Showed RTC captures a new dimension when compared to a set of existing coupling metrics
- Showed RTC is useful for impact analysis
- Showed combining RTC with other conceptual coupling metrics provides superior accuracy for impact analysis

Thank you. Questions?

SEMERU @ William and Mary

http://www.cs.wm.edu/semeru/