Meerkat

Community Evolution Prediction in Dynamic Social Networks

(A277)

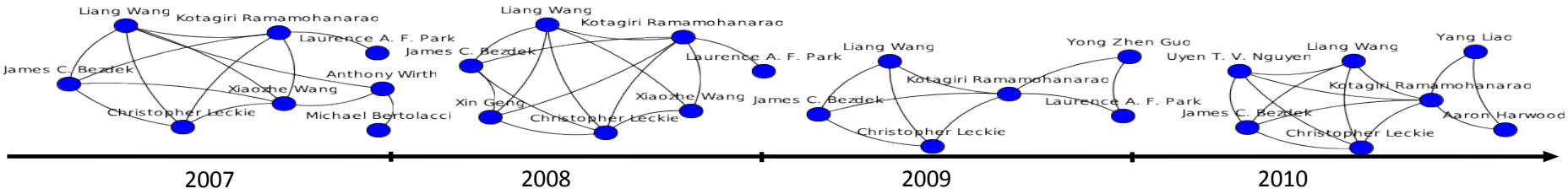
Mansoureh Takaffoli, Reihaneh Rabbany, Osmar R. Zaïane

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada

Overview

Predict future of dynamic networks has application in recommendation systems and customer targeting

- **Predicting community events and transitions**
  - Features related to community's structure, history, and influential members
  - High accuracy in experiments
  - Also detects predictive features per event



介紹

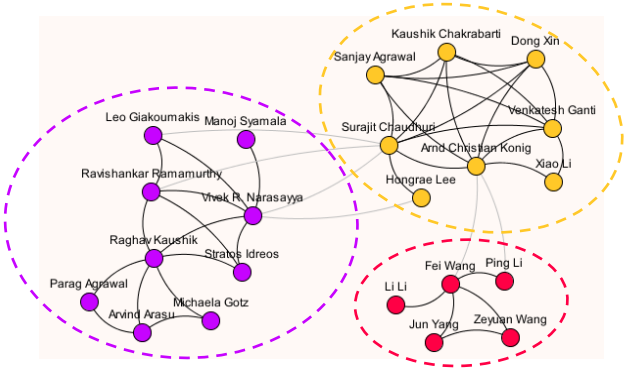
Introduction

Motivation

- Networks are dynamic
  - Relationships based on interactions are intrinsically temporal and changing over time
    - e.g. friendships, co-authorships, email
- **Modeling and Predicting evolution of a network**
  - Application in viral marketing [Leskovec et al. 2007], revenue maximization [Akhlaghpour et al. 2010], and social influence [Agarwal et al. 2008]
- Previous studies: mostly **macroscopic** graph structure, or **microscopic** properties from the point of view of a single node or edge

Community Level Prediction

- Predicting the trend of one **mesoscopic structure**, called community
  - Densely connected subset of nodes that are loosely connected to others



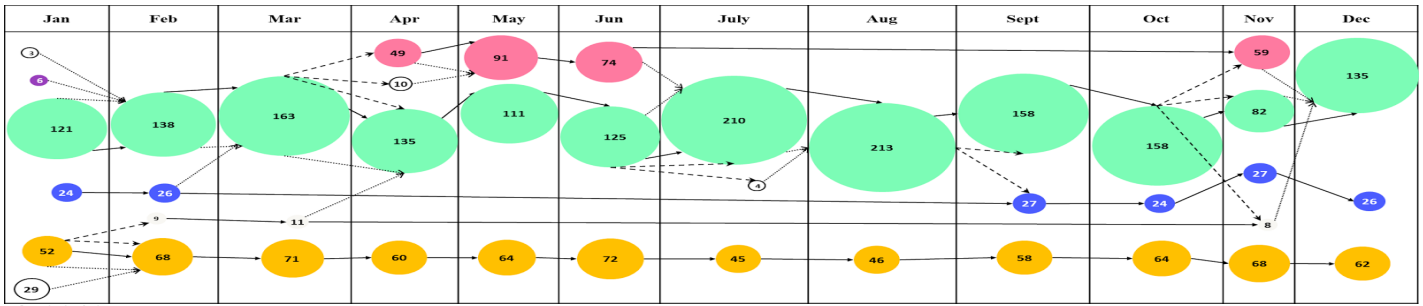
- Evolution of communities is related to important social phenomena such as homophily [McPherson et al. 2001] and influence [Anagnostopoulos et al. 2008]

Types of Communities

- Implicit v.s. Explicit communities
  - Explicit communities: defined based on rules
    - e.g. employees of a company, or students participating in a course
  - **Implicit communities**: defined and formed based on interactions
- Overlapping v.s. Non-Overlapping
  - Each individual belongs to only one community
    - Main engagement platform for the individual

Events and Transitions

- Community experiences different events and transitions during its life
  - An individual can move from one community and join another one
  - Amount of interactions between members of a community changes over time



Takoffoli et al. 2013

Predict Event and Transition

- Based on structural and temporal properties
  - Relation between **behaviour of individuals** and **future of their communities**
    - More central ⇒ more influential
  - Comprehensive predictive model for different events and transitions, and outlining predictive features
  - No consecutive snapshots

方法

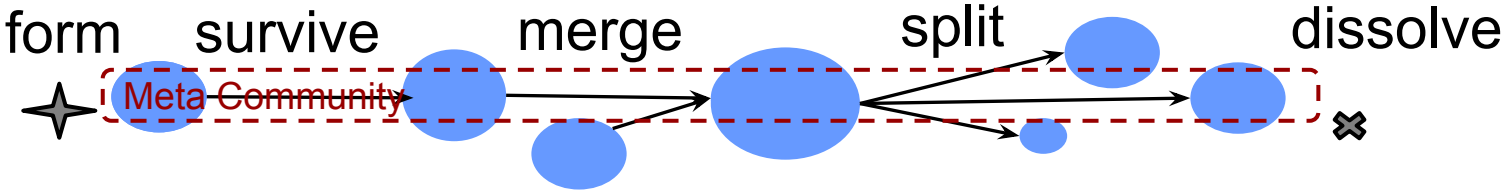
Method

Definitions (from Takaffoli et al. 2011)

- Similar communities
  - Ratio of mutual members > threshold

$$\mathrm{sim}(C_i^p, C_j^q) = \begin{cases} \frac{|V_i^p \cap V_j^q|}{\max(|V_i^p|, |V_j^q|)} & if \ \frac{|V_i^p \cap V_j^q|}{\max(|V_i^p|, |V_j^q|)} \geq k \\ 0 & otherwise \end{cases}$$

- 5 community events



- Meta Community

$$\begin{aligned} &(a) \quad 1 \leq t_1 < t_2 < \ldots < t_m \leq n, \\ &(b) \quad \forall\, t_i,\ t_1 < t_i \leq t_m\ \exists\, t_j < t_i :\ \mathrm{sim}(C_{t_j}^{p_j}, C_{t_i}^{p_i}) > 0 \end{aligned}$$

Problem Formulation

- Predictive models
  - Response variable: quantifies a particular change in a community ⇒ Community Events
  - Feature: influences the response variables ⇒ Properties of communities
    - Its influential members
    - Structural properties of the community
    - Temporal changes of features
    - Contextual attributes as features

Community Transitions and Events as Response Variables

- survive {true, false}
  - size {expand, shrink}
  - cohesion {cohesive, loose}
- merge {true, false}
- split {true, false}

- Binary
- Not exclusive ⇒ separate models

Properties of Community as Features

- Properties of its influential members
- Structural properties of the community
- Temporal changes of features
- Contextual attributes as features

Properties of its influential members

small set of individuals influence other members and shape the fate of community

- Influential nodes: Leaders by Abnar et al. (A328)
  - detected from pdf of closeness centrality
  - avg degree and closeness centrality

…

Also for outermosts: set of least significant individuals

Structural properties of the community

- Size
  - number of nodes
- Cohesion
  - how closely its members interact with each other relative to outside of the community
- Density
  - ratio of edges to the maximum possible edge
- Clustering coefficient
  - mean of clustering coefficient of all its members

Temporal & Contextual features

- Temporal changes of features
  - Current rate of change in each feature
    - e.g. ΔClosenessLeaders
  - Previous events and transitions
    - e.g. PreviousSurvive
- Contextual attributes as features
  - stable topics (most frequent keywords),
  - stable topics of leaders

changes in the topics discussed within a community or by its influential members affect its future

TABLE I. PROBLEM FORMULATION: FEATURES AND RESPONSE VARIABLES RELATED TO A COMMUNITY

Category	Feature	Description	Domain
Influential Member	ClosenessLeaders	average of closeness centrality of leaders	$(0, 1]$
	DegreeLeaders	average of degree centrality of leaders	$(0, 1]$
	LeadersRatio	ratio of leaders	$(0, 1]$
	OutermostsRatio	ratio of outermosts	$[0, 1]$
Community	Density	ratio of edges to maximum possible edges (Equation3)	$(0, 1]$
	ClusteringCoefficient	ratio of edges between neighbours of a nodes to maximum possible edges (Equation4)	$(0, 1]$
	NodesNumber	number of nodes	$[2, \infty)$
	Cohesion	ratio of members interact with each other to outside of the community (Equation 2)	$(0, \infty)$
	AverageCloseness	average of closeness centrality scores	$(0, 1]$
	VarianceCloseness	variance of closeness centrality scores	$[0, 1]$
	AverageDegree	average of degree centrality scores	$(0, 1]$
	VarianceDegree	variance of degree centrality scores	$[0, 1]$
Temporal	ΔClosenessLeaders	difference between average of closeness centrality of leaders	$(0, 1]$
	ΔDegreeLeaders	difference between average of degree centrality of leaders	$(0, 1]$
	ΔLeadersRatio	difference between ratio of leaders	$[0, 1]$
	ΔOutermostsRatio	difference between ratio of outermosts	$[0, 1]$
	ΔDensity	difference between density	$[0, 1]$
	ΔClusteringCoefficient	difference between clustering coefficient	$[0, 1]$
	ΔAverageCloseness	difference between average of closeness centrality scores	$[0, 1]$
	ΔVarianceCloseness	difference between variance of closeness centrality scores	$[0, 1]$
	ΔAverageDegree	difference between average of degree centrality scores	$[0, 1]$
	ΔVarianceDegree	difference between variance of degree centrality scores	$[0, 1]$
	JoinNodesRatio	percentage of nodes joining to this community	$[0, 1]$
	LeftNodesRatio	percentage of nodes leaving this community	$[0, 1]$
	Similarity	similarity between community and its previous instance (Equation 1)	$[k, 1]$
	LifeSpan	number of snapshots between this community and the first instance of the same community	$[1, n]$
	PreviousSurvive	survive event occurred for previous instance of the community	$\{\text{true}, \text{false}\}$
	PreviousMerge	merge event occurred for previous instance of the community	$\{\text{true}, \text{false}\}$
	PreviousSplit	split event occurred for previous instance of the community	$\{\text{true}, \text{false}\}$
	PreviousSizeTransition	size transition occurred for previous instance of the community	$\{\text{expand}, \text{shrink}\}$
	PreviousCohesionTransition	cohesion transition occurred for previous instance of the community	$\{\text{cohesive}, \text{loose}\}$
	StableTopics	stable topics between community and its previous instance	$\{\text{true}, \text{false}\}$
	StableLeaderTopics	stable topics between leaders of community and leaders of its previous instance	$\{\text{true}, \text{false}\}$
Response variable	survive	survive event occurred for the community	$\{\text{true}, \text{false}\}$
	merge	merge event occurred for the community	$\{\text{true}, \text{false}\}$
	split	split event occurred for the community	$\{\text{true}, \text{false}\}$
	size	size transition occurred for the community	$\{\text{expand}, \text{shrink}\}$
	cohesion	cohesion transition occurred for the community	$\{\text{cohesive}, \text{loose}\}$

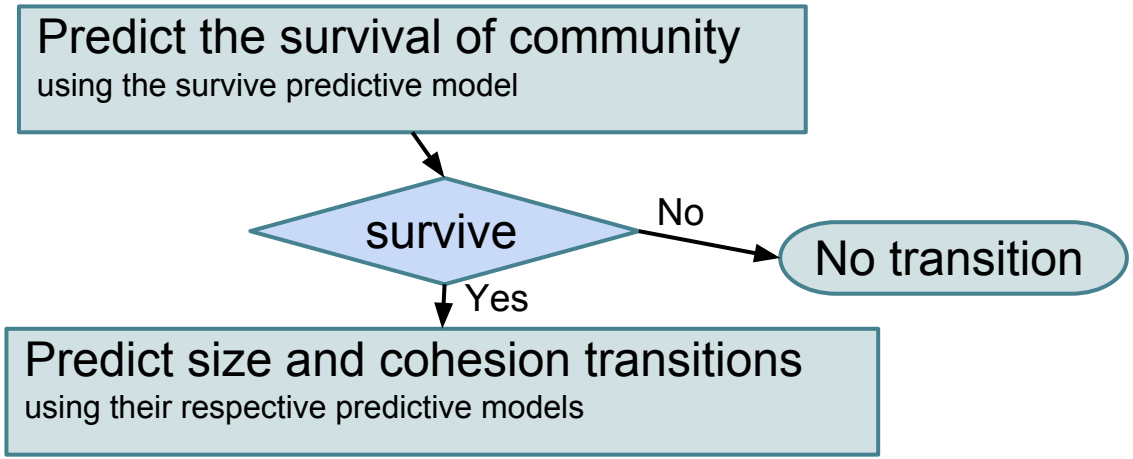
COMMUNITY EVOLUTION PREDICTION

Typical ML task

- logistic regression [LogitBoost]
  - most significant feature set by forward stepwise additive regression [Landwehr, et al. 2005],
- Naive Bayes classifier, Bagging classifier, Decision Table classifier, Decision Stump classifier, J48 Decision tree, Bayesian Networks classifier, Simple CART classifier, Support Vector Machine (SVM) classifier, and Neural network classifier [WEKA]
  - wrapper method: select features based on the error rate of each classifier

Two-stage cascade predictive model

information collected from the output of a first stage is used as additional information for the second stage in the cascade



结果

Results

Case Studies, Datasets

- **Enron email dataset**
  - Emails between employees of the Enron Corp.
  - year 2001: company declared bankruptcy
  - A total of **210** nodes
  - each month one snapshot
- **DBLP dataset**
  - co-authorship network related to field of DB and DM
  - from year 2001 to 2010
  - A total of **19461** authors
  - snapshot is one year

Case Studies, Datasets

- **Community detection**
  - in each snapshot by [Chen et al. 2009] algorithm
- **Topic extraction**
  - per community by KEA [Witten et al. 1999]
- **Community event detection**
  - by MODEC [Takaffol et al. 2011]
- **Event prediction model**
  - proposed feature set and 10-fold cross-validation

Results on Enron Email Dataset

113 community instances, 61 survive, 27 split, 55 merge

TABLE II. ENRON: SURVIVE EVENT PREDICTION

Event	Predictive Model	Accuracy	Precision	Recall	F-measure
Survive	**SVM**	**70**	0.7	0.7	0.7
	Bagging	70	0.7	0.7	0.7
	Decision Stump	68.333	0.686	0.683	0.683
	Naïve Bayes	67.5	0.675	0.675	0.675
	Neural Network	66.666	0.667	0.667	0.667
RSurvive[6]	**Decision Table**	**90.566**	0.911	0.90	0.905
	Neural Network	89	0.841	0.84	0.839
	SVM	87.735	0.879	0.877	0.877
	BayesNet	85.849	0.862	0.858	0.858
	Logistic Regression	83.962	0.84	0.84	0.84

Results on Enron Email Dataset

TABLE IV. ENRON: COMMUNITY TRANSITIONS PREDICTION

Transition	Predictive Model	Accuracy	Precision	Recall	F-measure
Size	**J48 Decision tree**	**73.684**	0.745	0.737	0.735
	Neural Network	70.175	0.716	0.702	0.698
	SVM	68.421	0.689	0.684	0.683
	Decision Stump	68.421	0.686	0.684	0.684
	Decision Table	68.421	0.686	0.684	0.684
Cohesion	**Decision Table**	**78.431**	0.796	0.784	0.783
	BayesNet	78.431	0.796	0.784	0.783
	Decision Stump	78.431	0.796	0.784	0.783
	J48 Decision tree	74.509	0.749	0.745	0.745
	Bagging	70.588	0.706	0.706	0.706

Results on Enron Email Dataset

TABLE III. ENRON: MERGE AND SPLIT EVENTS PREDICTION

Event	Predictive Model	Accuracy	Precision	Recall	F-measure
Split	**SVM**	**85.965**	0.86	0.86	0.86
	BayesNet	85.965	0.861	0.86	0.859
	Neural Network	84.21	0.843	0.842	0.842
	SimpleCART	83.626	0.837	0.836	0.836
	Decision Table	83.041	0.831	0.83	0.83
Merge	**Naïve Bayes**	**77.333**	0.779	0.773	0.773
	Neural Network	74.667	0.747	0.747	0.747
	Logistic Regression	72	0.72	0.72	0.72
	SVM	70.667	0.713	0.707	0.705
	BayesNet	68	0.736	0.68	0.662

Results on DBLP Database

1949 community instances, 1813 survive, 166 split, 306 merge

TABLE V. DBLP: SURVIVE EVENT PREDICTION

Event	Predictive Model	Accuracy	Precision	Recall	F-measure
Survive	**BayesNet**	**61.969**	0.62	0.62	0.62
	Naïve Bayes	61.583	0.618	0.616	0.614
	Logistic Regression	61.555	0.618	0.616	0.614
	Decision Table	60.921	0.609	0.609	0.609
	Bagging	60.811	0.609	0.608	0.607
RSurvive	**Decision Table**	**83.857**	0.878	0.839	0.834
	Decision Stump	83.857	0.878	0.839	0.834
	Neural Network	83.434	0.869	0.834	0.83
	BayesNet	82.164	0.845	0.822	0.819
	SimpleCART	81.257	0.855	0.813	0.807

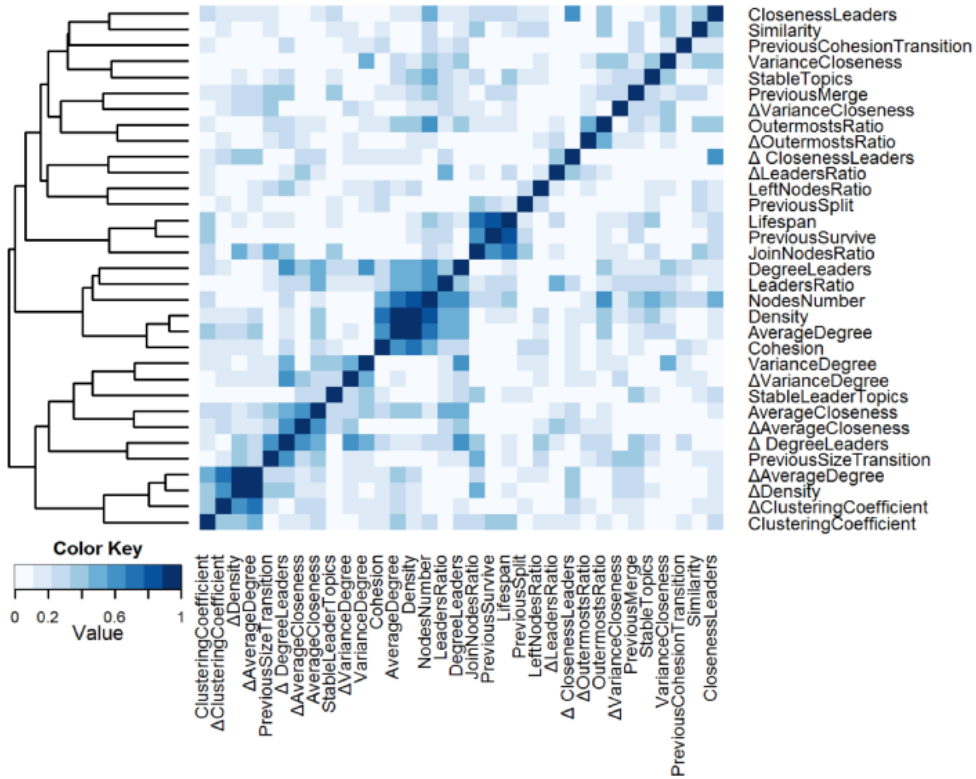
Results on DBLP Database

TABLE VI. DBLP: MERGE AND SPLIT EVENTS PREDICTION

Event	Predictive Model	Accuracy	Precision	Recall	F-measure
Split	**Naïve Bayes**	**80.723**	0.808	0.807	0.807
	SVM	80.723	0.807	0.807	0.807
	BayesNet	79.819	0.798	0.798	0.798
	Decision Stump	79.217	0.792	0.792	0.792
	Bagging	78.916	0.789	0.789	0.789
Merge	**Naïve Bayes**	**62.582**	0.626	0.626	0.625
	SimpleCART	61.928	0.62	0.619	0.619
	Decision Table	60.621	0.607	0.606	0.605
	Logistic Regression	59.967	0.602	0.6	0.597
	Bagging	59.967	0.6	0.6	0.6

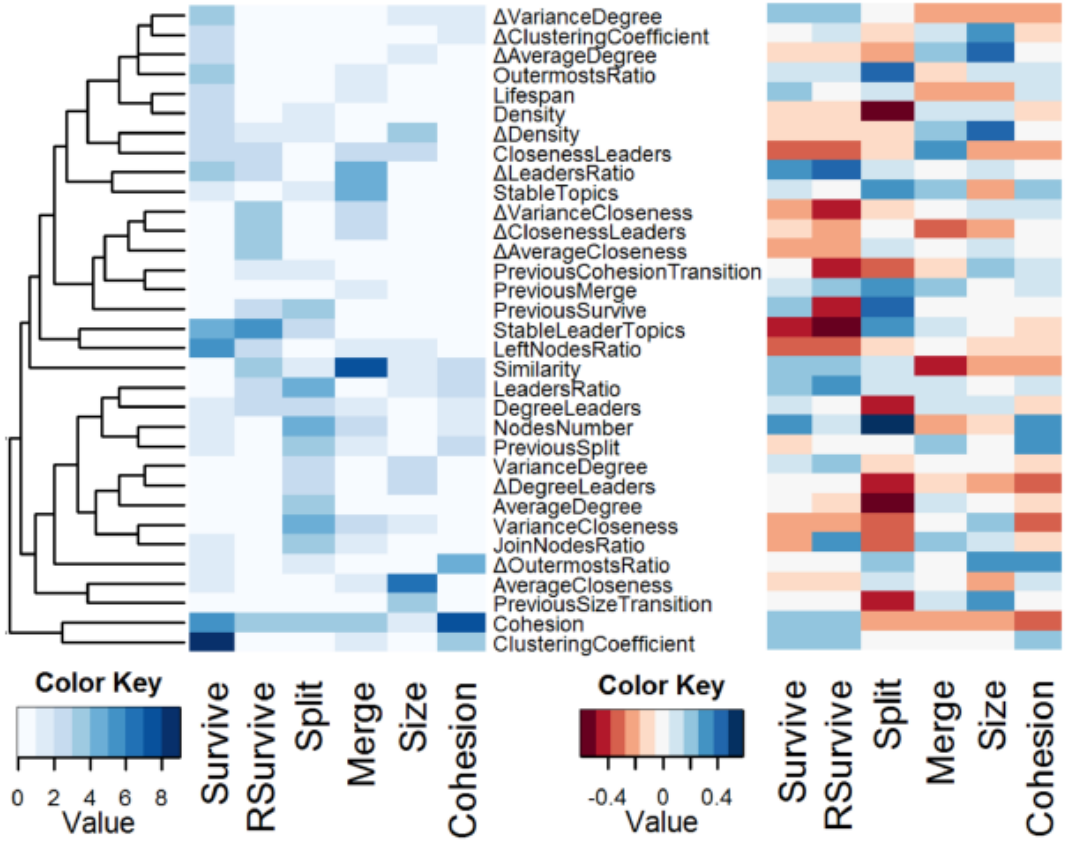
Correlation between Features

- Density, Clustering-Coefficient, AverageDegree, and AverageCloseness features are correlated
- Non redundant features



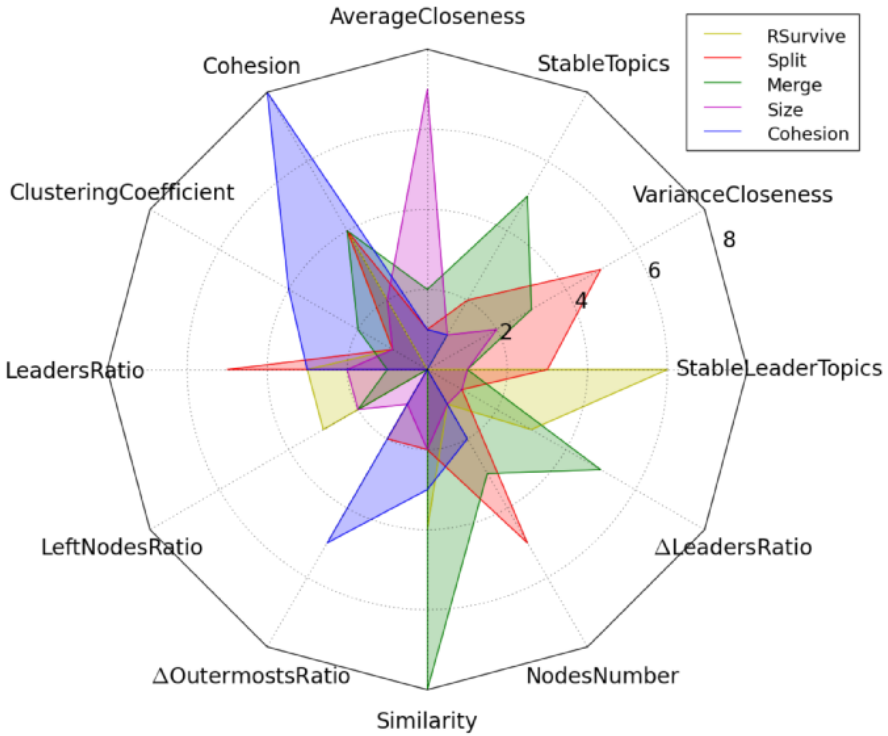
Ensemble Analysis

- Survival
  - ClusteringCoefficient
  - Cohesion
  - LeftNodesRatio
- Split
  - LeadersRatio
  - NodesNumber
- Merge
  - Similarity

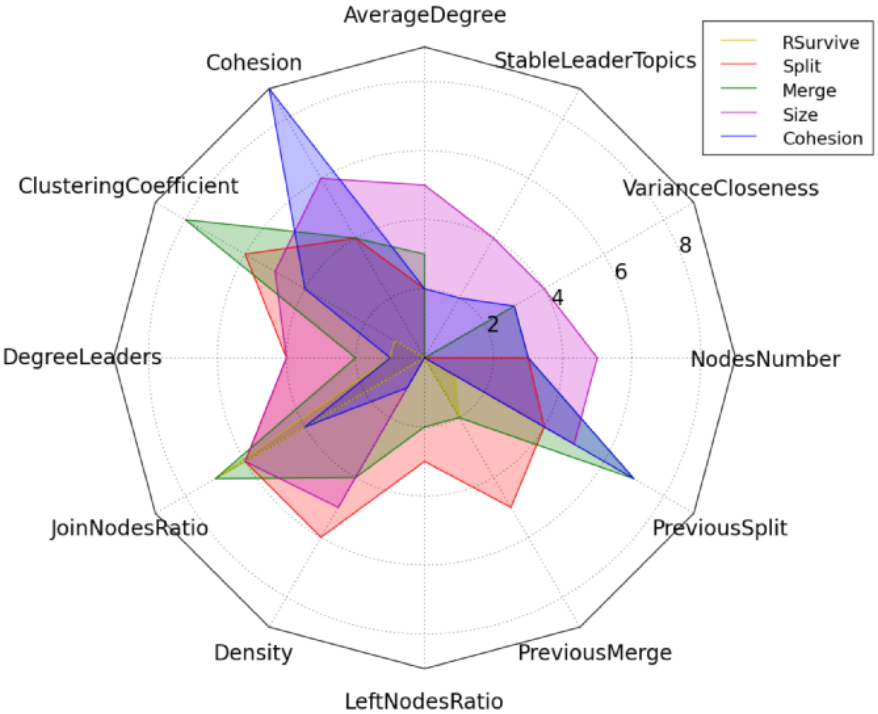


Comparison of prominent features

ENRON

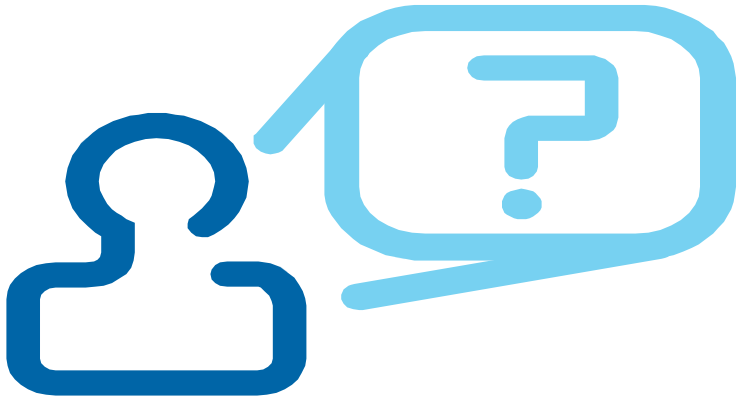


DBLP



Thank You

谢谢



Extra

Cohesion Formula

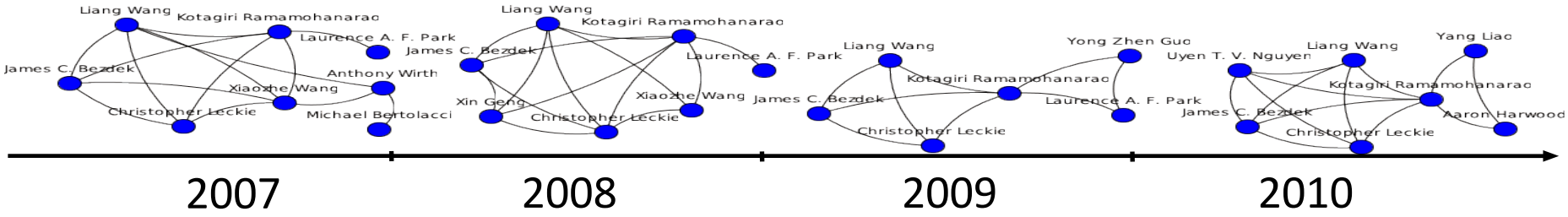
how closely its members interact with each other relative to outside of the community

Cohesion: Cohesion of a community C_i^p at snapshot i is:

$$\text{cohesion}(C_i^p) = \frac{\frac{2|E_i^p|}{|V_i^p|(|V_i^p|-1)}}{\frac{|OE_i^p|}{|V_i^p|(|V_i|-|V_i^p|)}} = \frac{2|E_i^p|(|V_i| - |V_i^p|)}{|OE_i^p|(|V_i^p| - 1)} \quad (2)$$

Dynamic Social Network Analysis

- Model network using time series graphs
- Characterize evolution of communities and entities



Goal: Detecting the evolution of communities in a dynamic network

MODEC Framework (CASON 2011)

- Critical events are used to characterize the evolution of communities.