A METHOD FOR CHARACTERIZING COMMUNITIES IN DYNAMIC ATTRIBUTED COMPLEX NETWORKS

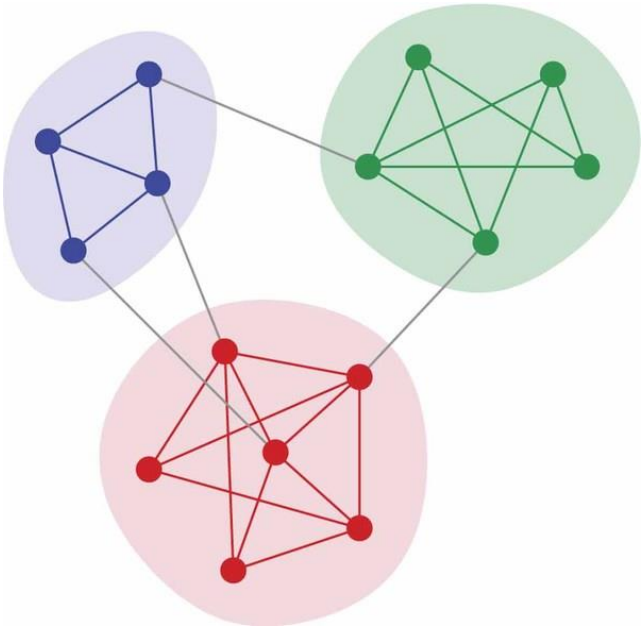
Günce Keziban Orman, Vincent Labatut, Marc Plantevit,
& Jean-François Boulicaut

INSA de Lyon - LIRIS
Galatasaray University – Computer Engineering Department

Plan

1. Introduction
   1. Community Detection
   2. Community Characterization (Novel Problem)
2. Problem Definition
   1. Sequence Mining
   2. Representing Dynamic Attributed Networks
3. Method
   1. Community Detection
   2. Node Identification
   3. Mining Closed Frequent Emerging Sequences
   4. Selecting Characterizing Patterns and Finding Outliers
4. Empirical Results
5. Conclusion

1.1. Community Detection



A very active sector concentrated on detection task

What about community interpretation?

1.2. Community Characterization Novel Problem

- Community = group of nodes with common properties
  - Node description
    - Topology (via network)
      - Relational Information
    - Attributes
      - Individual Information
- Community = group of nodes with common behaviors
  - Temporal evolution of the descriptions

Interpreting communities in a systematic way by considering common changes on the node descriptions over the time

2. Problem Definition

- Identifying the discriminant features of the nodes of each community
- 1. What are the features?
  - Encoding all the information describing the evolution of each node from each community in a complete and compact way
- 2. How to measure their typicality?
  - Choosing objective criteria to measure the power of representation of each sequence
    - Informative
    - Prevalent
    - Distinctive

2.1. Sequence Mining

- Finding correlated elements in an ordered set of elements
- Given a set of sequences, finding the complete set of **frequent sub-sequences**

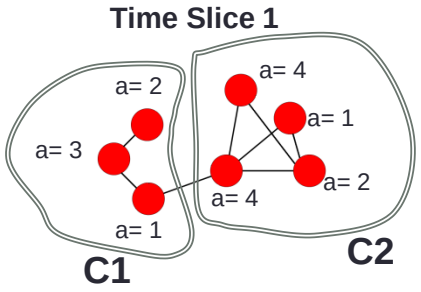
A *sequence database*

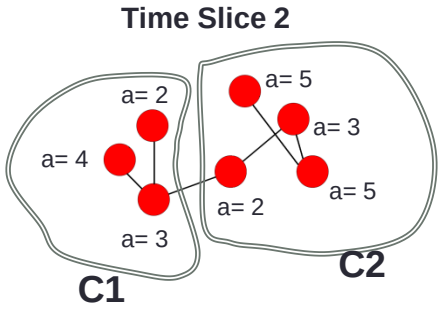
SID	Sequences
10	<a(**ab**c)(a**c**)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(**ab**)(df)**c**b>
40	<eg(af)cbc>

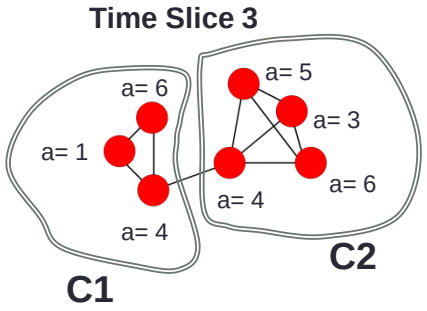
- A *sequence* <(ef)(ab)(df)cb> is an ordered set of itemsets.
- α = <(ab)c> is a *sub-sequence* of β= <a(**ab**c)(a**c**)d(cf)>

- Given *support threshold* $\min_{sup}$ = 2/4, <(ab)c> is a *frequent sequence*
- A sequence is *closed* if there is no other super-sequence of it with the same support

2.2. Representing Dynamic Attributed Networks







Items: Node descriptors with their values
{a=1,a=2,a=3,a=4,a=5,a=6}

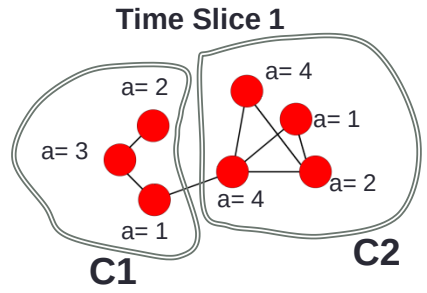
Node Sequence: *u(n1)=<(a=2)(a=2)(a=6)>*

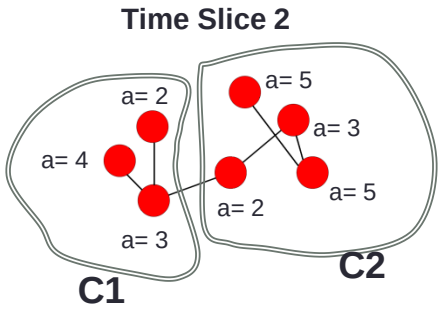
Node Sequence concatenation:
u(n1) ●C(n1) = <(a=2)(a=2)(a=6)(C1)>

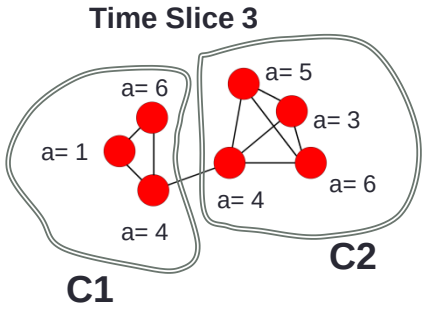
Node Sequence Database

N1 <(a=2)(a=2)(a=6)(C1)>
N2 <(a=3)(a=4)(a=1)(C1)>
N3 <(a=1)(a=3)(a=4)(C1)>
N4 <(a=4)(a=2)(a=4)(C2)>
N5 <(a=2)(a=5)(a=6)(C2)>
N6 <(a=1)(a=3)(a=3)(C2)>
N7 <(a=4)(a=5)(a=5)(C2)>

2.2. Representing Dynamic Attributed Networks







Choosing Criteria

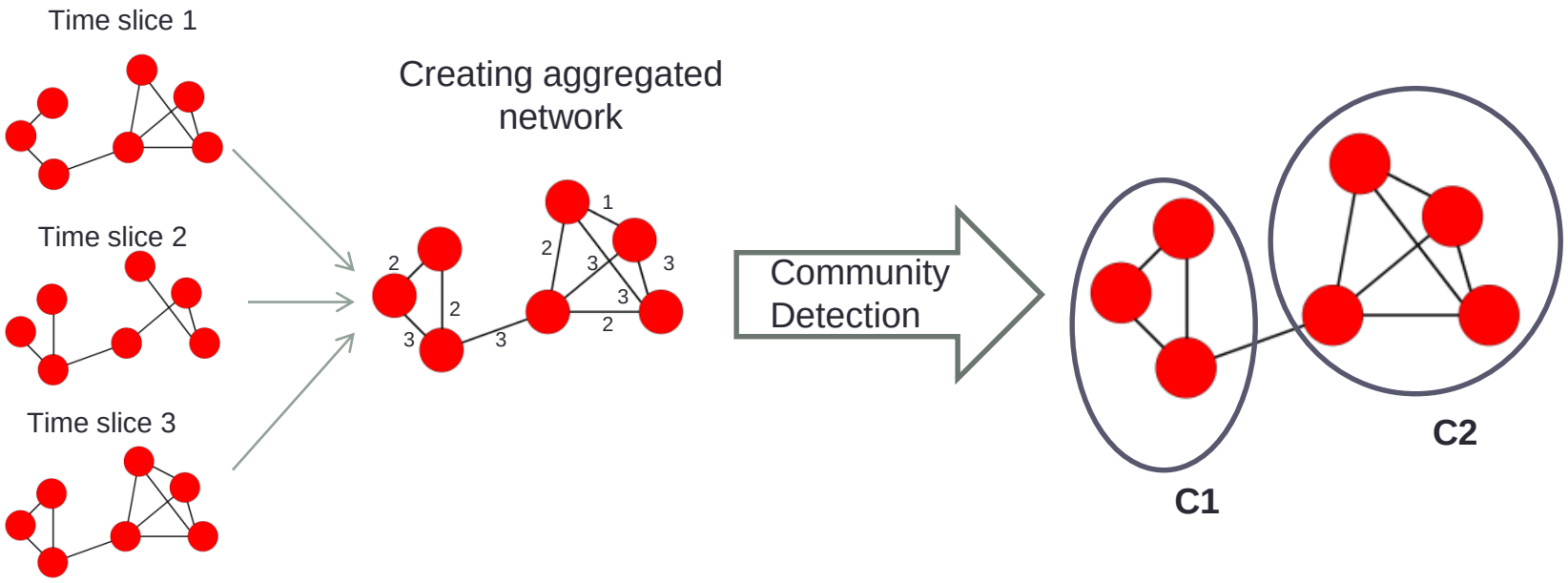
- **Prevalent:** Frequent sequences for min_{sup}
- **Informative:** Closed sequences
- What about **distinctiveness**?
  - Emerging Sequences
  - Growth rate of s relatively to a community C is
    - Gr(s,C) = sup(s,C)/(sup(s)-sup(s,C))
    - e.g., Gr(<(a=2)(a=2)>,C1)= ∞

3. Method

- **Step1.** Community Detection
- **Step2.** Mining Closed Frequent Emerging sequences
- **Step3.** Selecting characteristic patterns and finding outliers

3.1. Community Detection

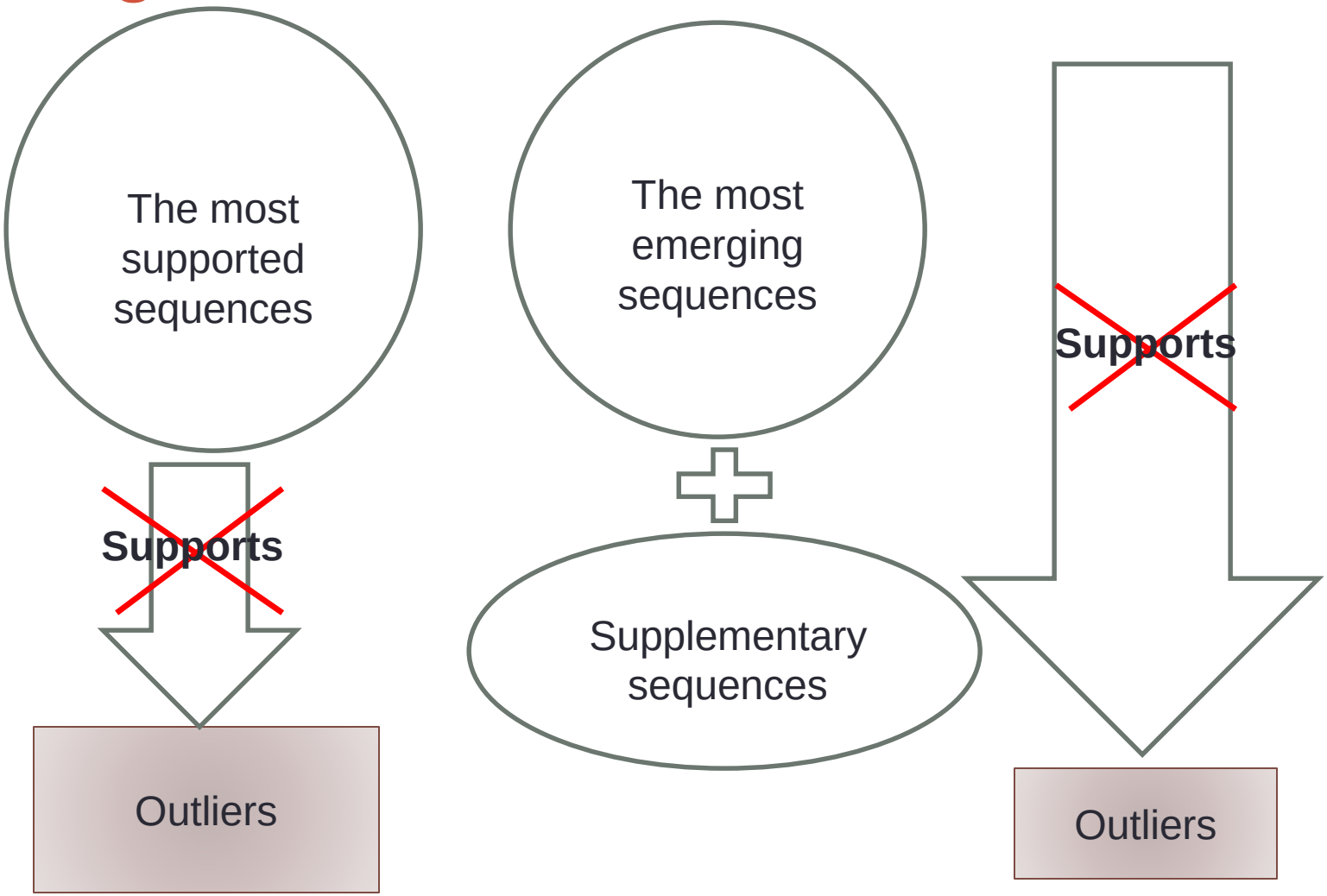
- Creating aggregated network
- Detecting communities by considering link weights
  - Louvain Method [Blondel *et al. 2008, JSTAT Mech*]



3.2. Mining Closed Frequent Emerging Sequences

- Creating sequence database
  - Attributes
  - 5 topological measures
    - *internal degree*, *local transitivity*, *within module degree*, *participation coefficient* and *embeddedness*
- Mining Closed Frequent Sequences
  - CloSpan [Yan *et al*. 2003,SDM] with $\min_{sup}$
- Calculating emergence
  - PostProcessing [Plantevit and Cremilleux 2009, IDA]

3.3. Selecting characteristic patterns and finding outliers



4. Empirical Results

- DBLP Dynamic Attributed Network
  - Nodes=researchers
  - Links=co-authorship
  - 10 time slice from 1990 to 2012
  - Attributes :
    - Publication number in 43 journal/conferences
    - Total conference publication number
    - Total journal publication number

- Focus : Communities containing >40 nodes

4. Empirical Results
I- The most supported

- The majority of the nodes of each community has a non-hub community role

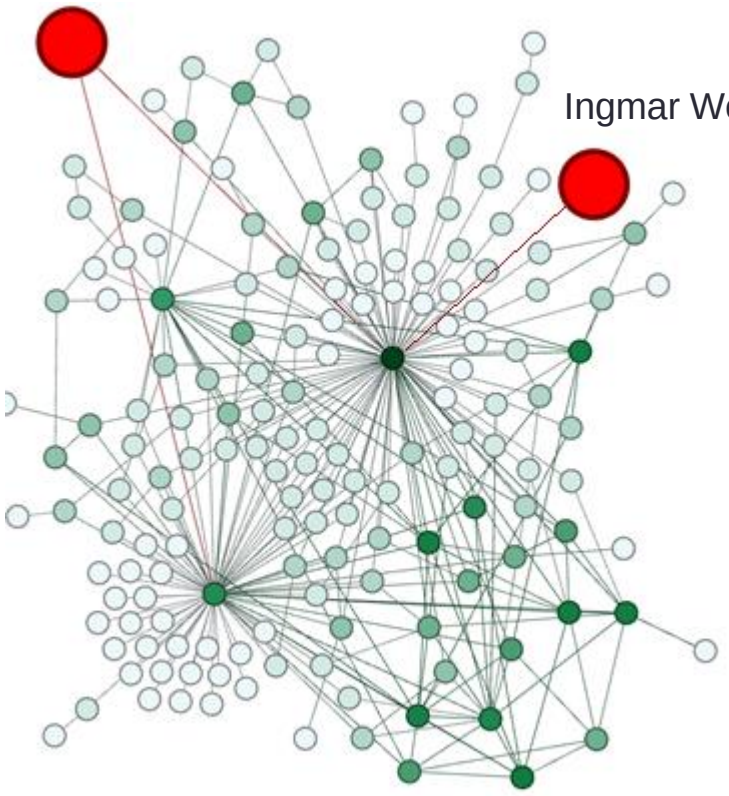
MOST SUPPORTED SEQUENCE SIZE FOR EACH COMMUNITY

Community ID	Community Size	Sequence Size	Support Value
38	335	2	0.99
40	43	8	0.97
77	41	7	1.00
115	125	1	1.00

Philip S. Yu,Jiawei Han,Beng C. Ooi

Hans-Peter Kriegel

4. Empirical Results
II- The most emerging



Representation of Community 45

<(*VLDB publication number=3*)
(*having few connections* **and** *community non-hub*)>
Gr. rate = 6.40 et sup=0.30

<(*community non-hub* **and** *total conference publication number between 1 and 5*) (*community non-hub, embedded to its community* **and** *ICDE publication number =3*)>
Gr. rate = 2.30 et sup=0.30

4. Empirical Results

- Both topology and attributes support the interpretation process
  - Main theme of communities
  - Relations of nodes
- Three types of outliers
  - Different theme
  - Changing theme
  - Rising actors

5. Conclusion & Perspectives

- Novel problem : community characterization
- Considering dynamic community detection
- Considering constraints on sequential patterns
  - Patterns including information both on topology and on attributes
- Application on different networks from different domains

Questions

Discrétisation

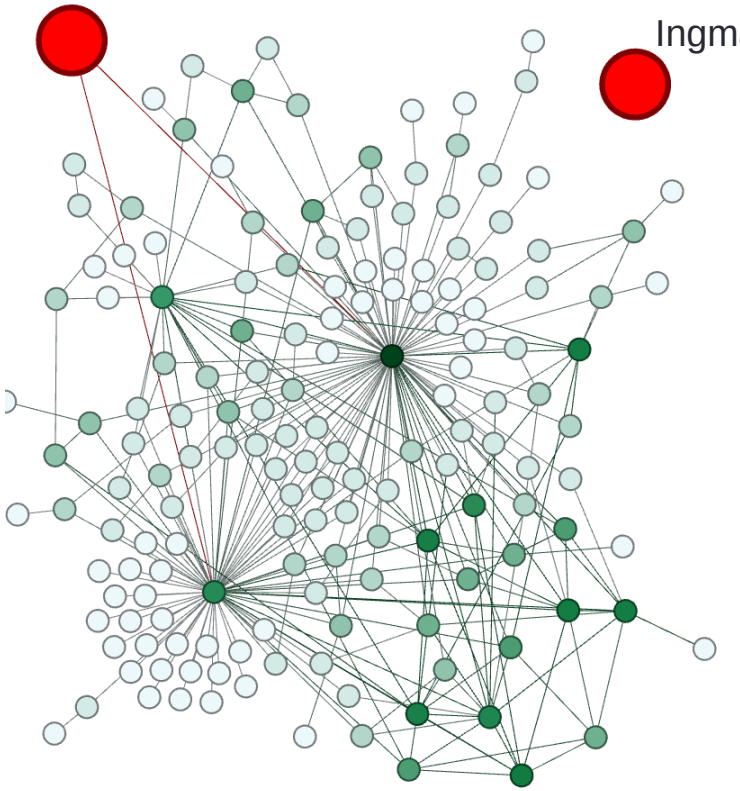
Descripteurs	**Ranges**				
Dégrée	[0;3] *Non populaire*	]3;10] *Peu populaire*	]10;30] *Populaire*	]30; ∞[*Très populaire*	
Transitivité	[0;0.35]	]0.35;0.5]	]0.5;0.7]	]0.7;1.0]	
Dégrée interne normalisé	]-∞;2.5] *non-hub communautaire*	]2.5, ∞[*Hub communautaire*			
Coefficient de Participation	[0;0.05] *Ultra Périphaire*	]0.05;0.6] *Périphaire*	]0.6;0.8] *Connecteur*	]0.8;1.0] *Kinless*	
Enchaissement	[0;0.3] *Peu appartenance*	]0.3;0.7] *Appartenance*	]0.7;1.0] *Très haute appartenance*		
Conf./Journal Pub.	1	2	3	4	[5; ∞[
Total Conf./Journal Pub.	[0;5[	[5;10[	[10;20[	[20;50[	[50; ∞[

3. Résultats

II- Les plus emergents

Anastasia Ailamaki

Ingmar Weber



Représentation de la communauté 45 p.r.a graph global pondère

<(*nombre de publication VLDB=3*) (*avoir peu de voisins* **et** *être non-hub communautaire*)>

Gr. rate = 6.40 et sup=0.30

<(*être non-hub communautaire* **et** *nombre de publication conférence total entre 1 et 5*) (*être non-hub communautaire*, *bien appartenir a sa communauté* **et** *nombre de publication ICDE =3*)>

Gr. rate = 2.30 et sup=0.30