Aalto University
School of Science

Active network alignment

Aristides Gionis

Department of Computer Science, Aalto University

users.ics.aalto.fi/gionis

Nov 23, 2017

data mining group in Aalto University
research themes

theoretical and application-driven research in data mining

most work focuses on graph mining

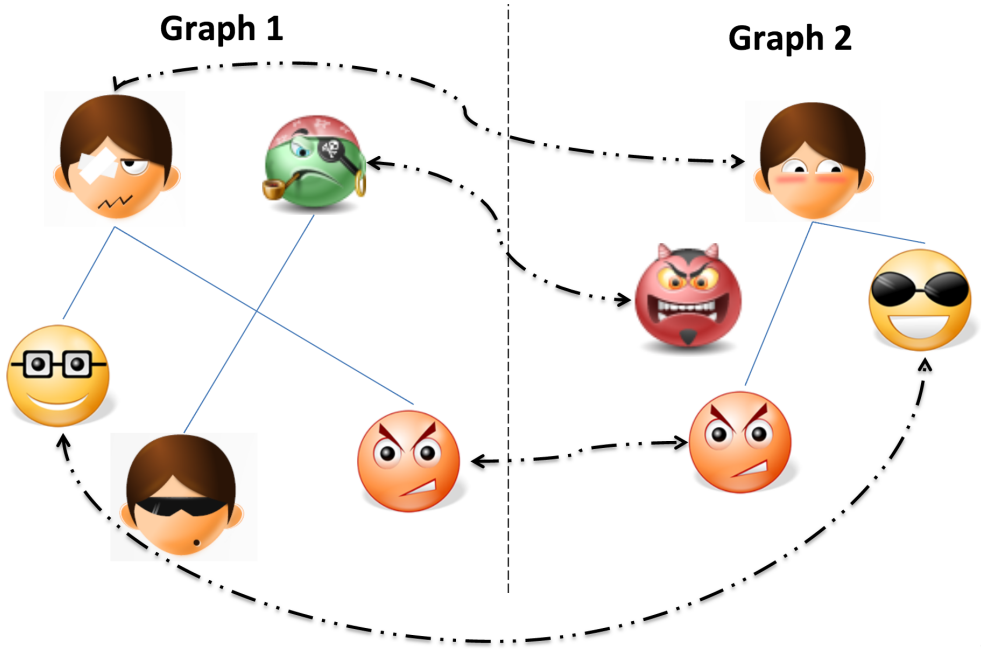
- finding dense subgraphs
- mining labeled and temporal networks
- network alignment
- network inference
- information propagation and opinion formation
- team formation

applications

- mining social media, e.g., studying polarization
- urban informatics, e.g., characterizing city neighborhoods

the network-alignment problem

- also known as: graph reconciliation, graph matching, collective entity resolution, etc.

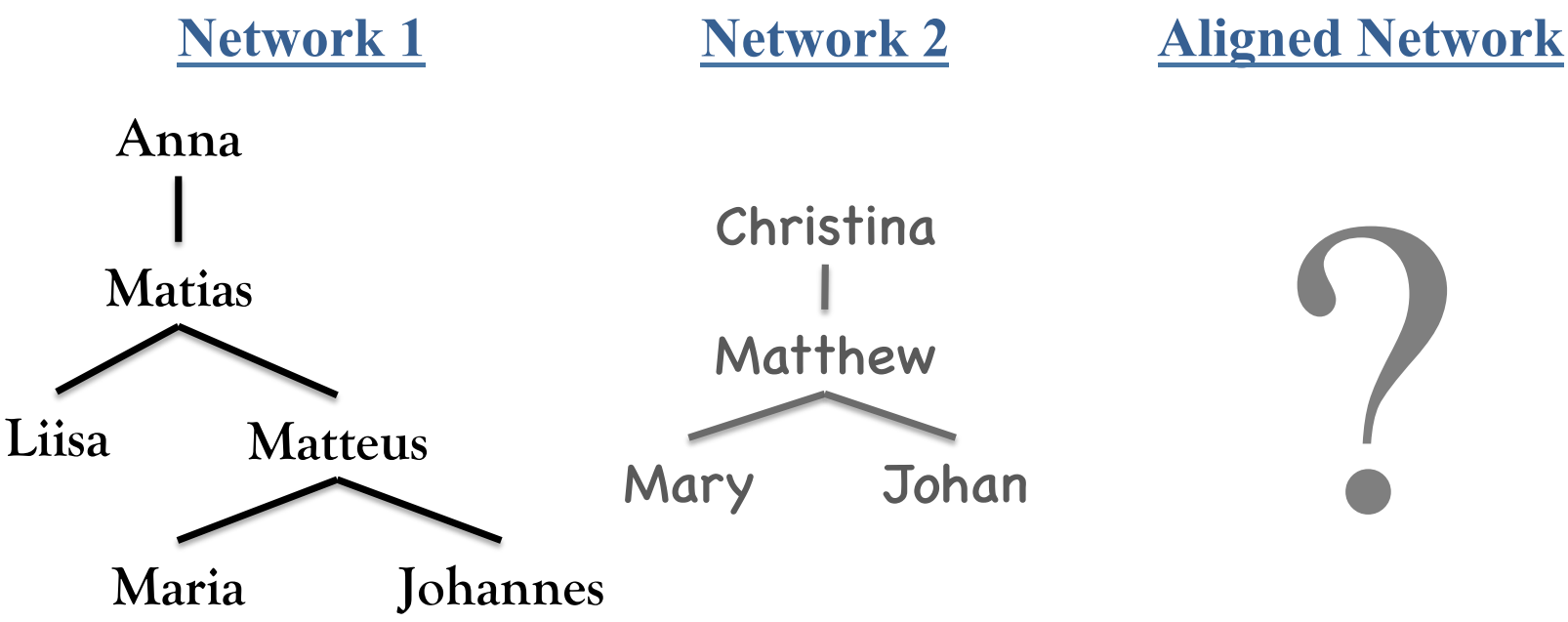


source: http://www.umiacs.umd.edu/~getoor/Tutorials/ER_KDD2013.pdf

- task: align nodes with similar attributes and similar neighbors

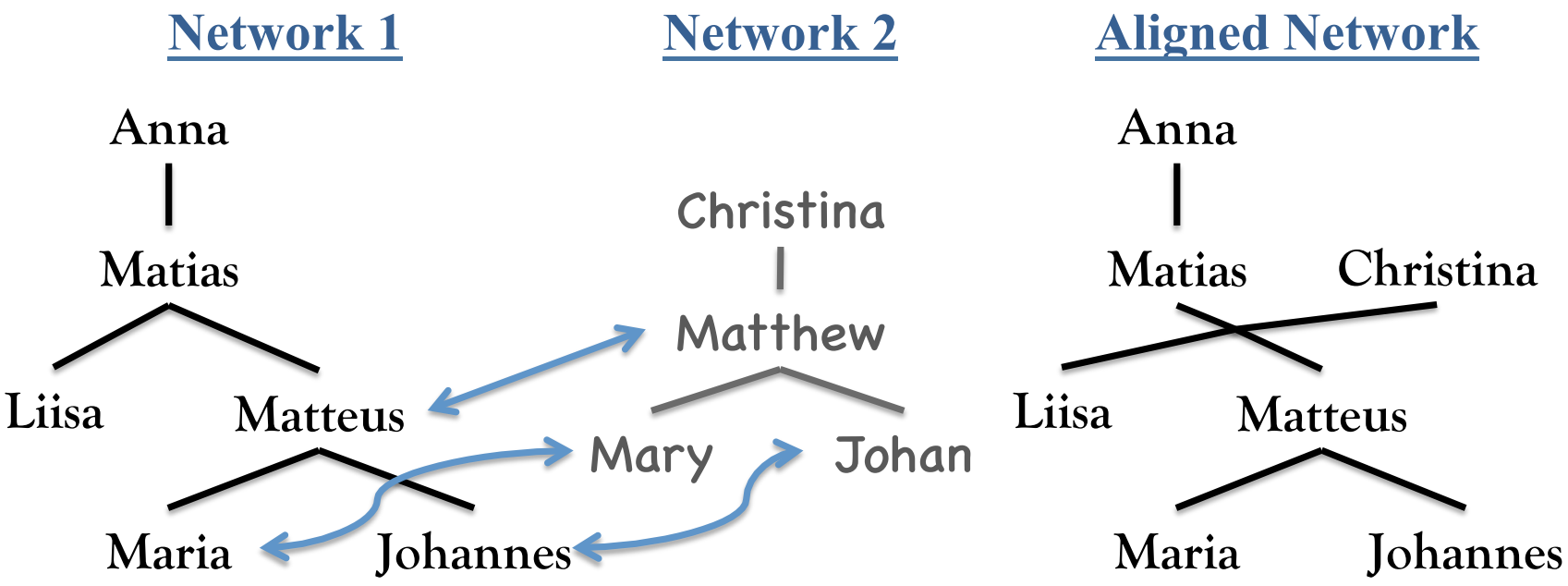
family trees

- input: family trees provided by different researchers
- output: a single, aligned family tree



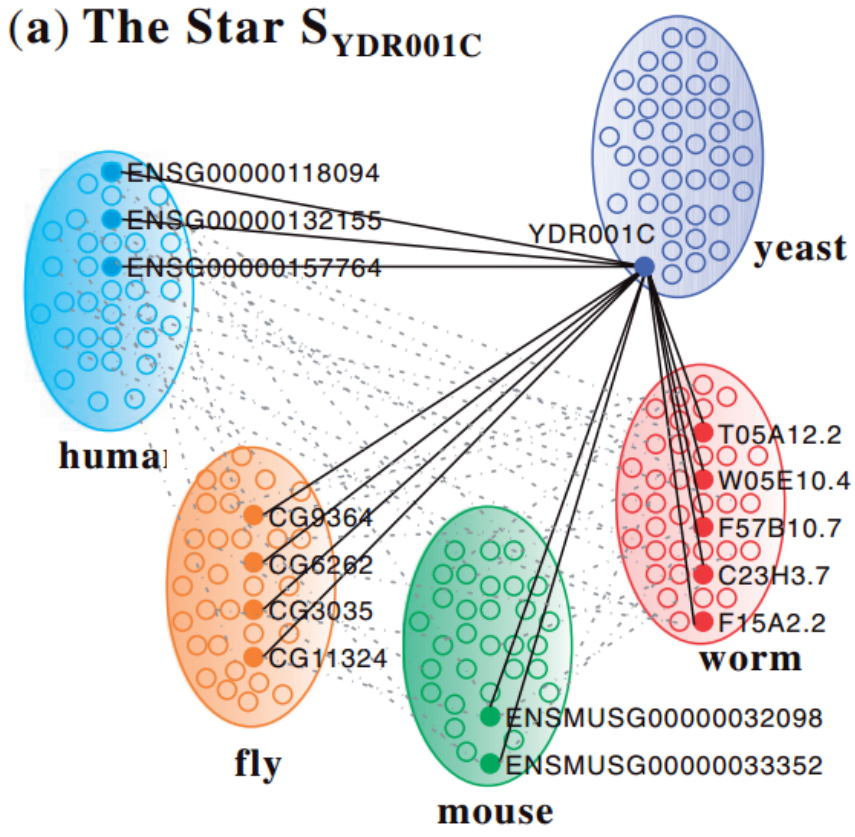
family trees

- input: family trees provided by different researchers
- output: a single, aligned family tree



protein-protein interaction networks

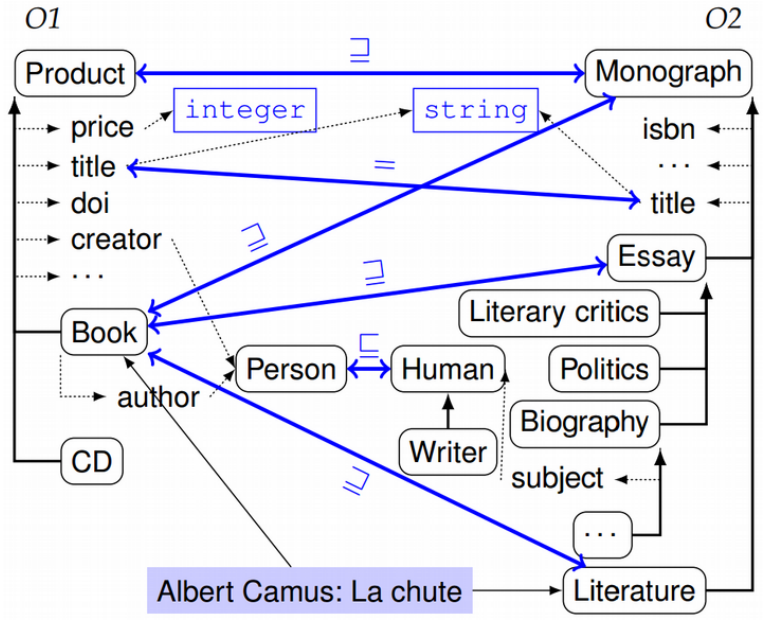
- task: align PPI networks of different species to identify functional orthologs across species



Source: Lia et al. Bioinformatics 2009

ontology matching

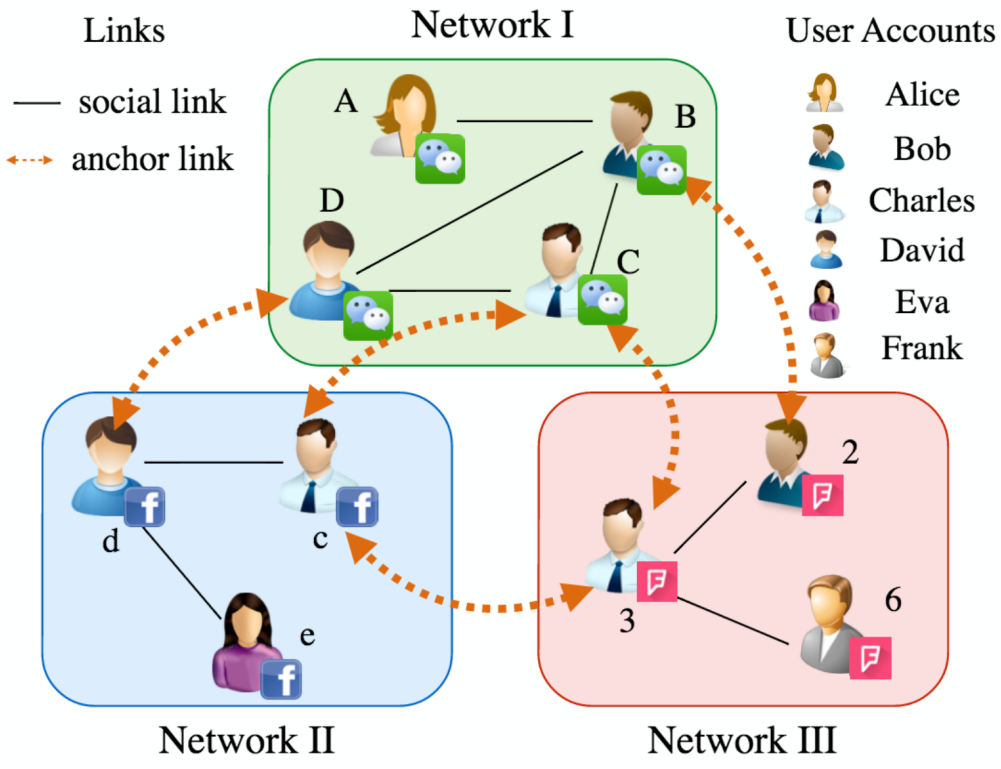
- a well-studied problem
- the 11th International Workshop on Ontology Matching organized in 2016
- applications in ontology evolution, data integration, data warehouses, etc.



Source: Shvaiko & Euzenat. Ontology matching: state of the art and future challenges. TKDE, 2013

social networks

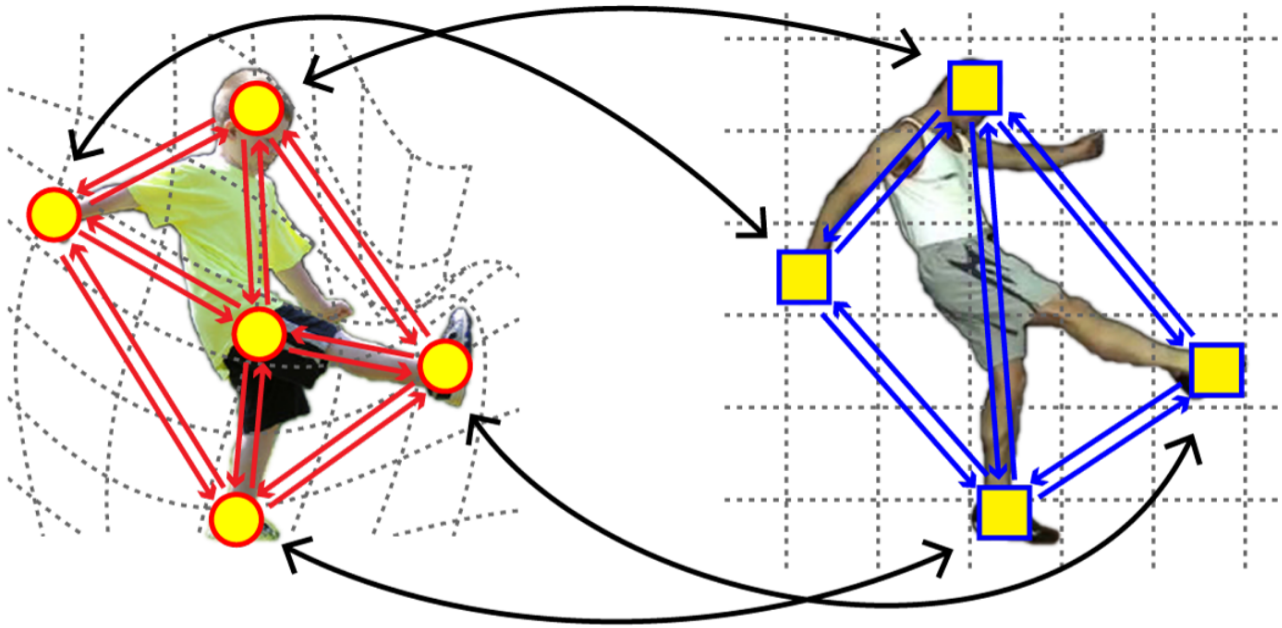
- task: find matching user profiles across services
- application: friend suggestions



Source: Zhang & Yu. ICDM 2015

computer vision and pattern recognition

- task: identify objects in a database
- an object is represented by a graph of landmarks containing information about features and spatial relationships



Source: http://www.f-zhou.com/gm.html

problem complexity

- graph isomorphism: not known to be in **P** nor **NP**-complete
- subgraph isomorphism: standard **NP**-complete problem

methods

- large number of practical methods

  notable methods:

  – IsoRank (linear system) [Singh et al., RECOMB 2007]

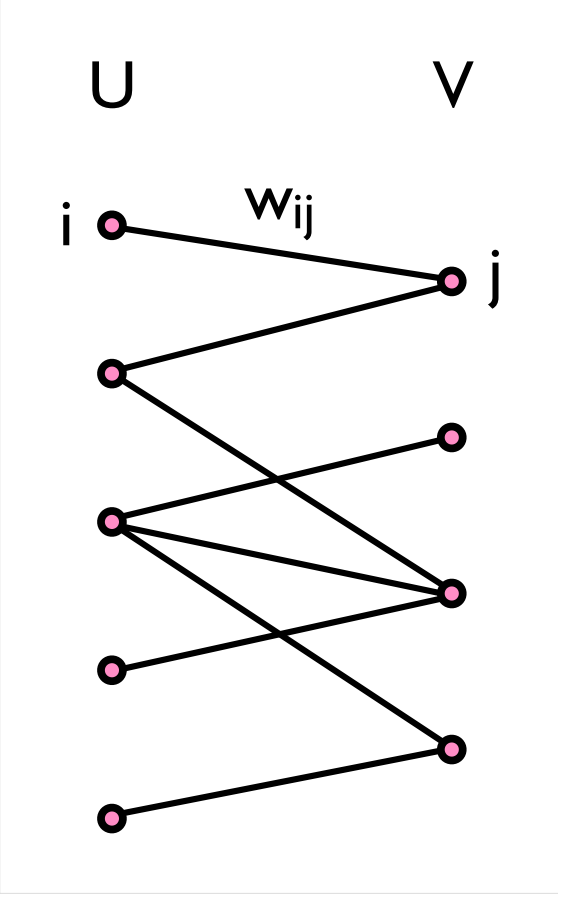
  – Natalie (global method) [Klau, BMC Bioinformatics 2009]

  – NetAlignMP (message passing) [Bayati et al., TKDD 2013]

  – UserMatch (local method) [Korula & Lattanzi, VLDB 2014]

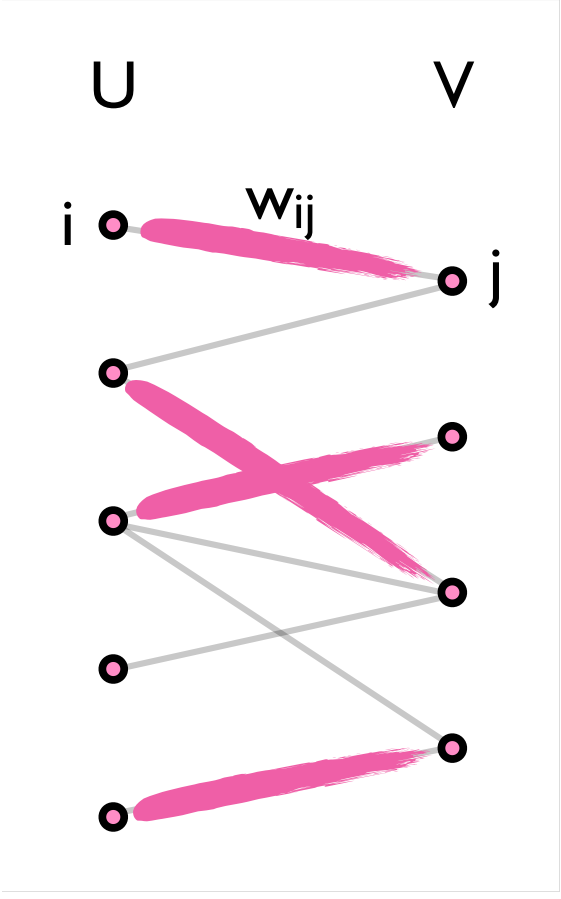
the matching problem

- given a bipartite graph $G = (U, V, E)$ with edge weights w_{ij}
- find a matching $M \subseteq E$ (each vertex is incident to at most one edge in M) that maximizes $\sum_{(i,j)\in M} w_{ij}$



the matching problem

- given a bipartite graph $G = (U, V, E)$ with edge weights w_{ij}
- find a matching $M \subseteq E$ (each vertex is incident to at most one edge in M) that maximizes $\sum_{(i,j)\in M} w_{ij}$
- problem solvable in polynomial time
  - Hungarian algorithm $\mathcal{O}(n^3)$



Natalie

- Gunnar Klau, *"A new graph-based method for pairwise global network alignment"*, BMC Bioinformatics, 2009
- a state-of-the-art network alignment method according to several independent studies

overview

1. formulate network alignment as a quadratic integer program
2. linearize the problem
3. apply Lagrangian relaxation to solve the linear problem

quadratic integer program

- input graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$
- binary x_{ij} indicates whether $i \in V_1$ is aligned with $j \in V_2$
- A^1 and A^2 are the adjacency matrices of G_1 and G_2
- $\sigma(i,j)$ is the similarity between node attributes, and
- g is a parameter

$$\max_x \quad \sum_{(i,j)\in V_1\times V_2} \sigma(i,j)x_{ij} + g \sum_{(i,j)\in V_1\times V_2} \sum_{(k,\ell)\in V_1\times V_2} A^1_{ik}A^2_{j\ell}x_{ij}x_{k\ell},$$

$$\text{such that} \quad \sum_j x_{ij} = 1,$$

$$\sum_i x_{ij} = 1,$$

$$x_{ij} \in \{0,1\}.$$

quadratic integer program

- input graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$
- binary x_{ij} indicates whether $i \in V_1$ is aligned with $j \in V_2$
- A^1 and A^2 are the adjacency matrices of G_1 and G_2
- $\sigma(i,j)$ is the similarity between node attributes, and
- g is a parameter

$$\max_x \quad \sum_{(i,j)\in V_1\times V_2} \sigma(i,j)x_{ij} + g \sum_{(i,j)\in V_1\times V_2} \sum_{(k,\ell)\in V_1\times V_2} A^1_{ik}A^2_{j\ell}x_{ij}x_{k\ell},$$

$$\text{such that} \quad \sum_j x_{ij} = 1,$$

$$\sum_i x_{ij} = 1,$$

$$x_{ij} \in \{0,1\}.$$

- **NP**-hard problem

linear integer program formulation

- replace $w_{ijk\ell} \leftarrow x_{ij}x_{k\ell}$ (guaranteed by constraints (1) and (2))

$$\max_{x,w} \quad \sum_{(i,j)\in V_1\times V_2} \sigma(i,j)x_{ij} + g \sum_{i,j\,\square\,k,\ell} w_{ijk\ell}$$

$$\text{such that} \quad \sum_j x_{ij} = 1, \quad \sum_i x_{ij} = 1,$$

$$\sum_\ell w_{ijk\ell} \le x_{ij}, \quad \sum_k w_{ijk\ell} \le x_{ij}, \tag{1}$$

$$w_{ijk\ell} = w_{k\ell ij}, \tag{2}$$

$$x_{ij}, w_{ijk\ell} \in \{0,1\}.$$

Lagrangian relaxation

- relax the symmetry constraint

$$Z_{LD}(\lambda) = \max_{x,w} \sum_{(i,j)\in V_1\times V_2} \sigma(i,j)x_{ij} + g \sum_{i,j\,\square\,k,\ell} w_{ijk\ell} + \sum_{i,j\,\square\,k,\ell} \lambda_{ijk\ell}(w_{ijk\ell} - w_{k\ell ij})$$

$$\text{such that} \quad \sum_j x_{ij} = 1, \quad \sum_i x_{ij} = 1,$$

$$\sum_\ell w_{ijk\ell} \le x_{ij}, \quad \sum_k w_{ijk\ell} \le x_{ij},$$

$$x_{ij}, w_{ijk\ell} \in \{0,1\}.$$

solving for a given λ

- problem can be written as

$$Z_{LD}(\lambda) = \max_{x} \quad \sum_{(i,j)} [\sigma(i,j) + v_{ij}(\lambda)]\, x_{ij}$$

$$\text{such that} \quad \sum_{j} x_{ij} = 1, \quad \sum_{i} x_{ij} = 1,$$

$$x_{ij} \in \{0,1\} \text{ for all } i,j.$$

- where

$$v_{ij}(\lambda) = \max_{w} \quad \sum_{(k,\ell):(i,j\,\square\,k,\ell)} [g + c_{ijk\ell}\lambda_{ijk\ell}]\, w_{ijk\ell}$$

$$\text{such that} \quad \sum_{\ell} w_{ijk\ell} \leq 1, \quad \sum_{k} w_{ijk\ell} \leq 1,$$

$$w_{ijkl} \in \{0,1\} \text{ for all } k,\ell.$$

- a total of $|V_1||V_2| + 1$ maximum matching instances

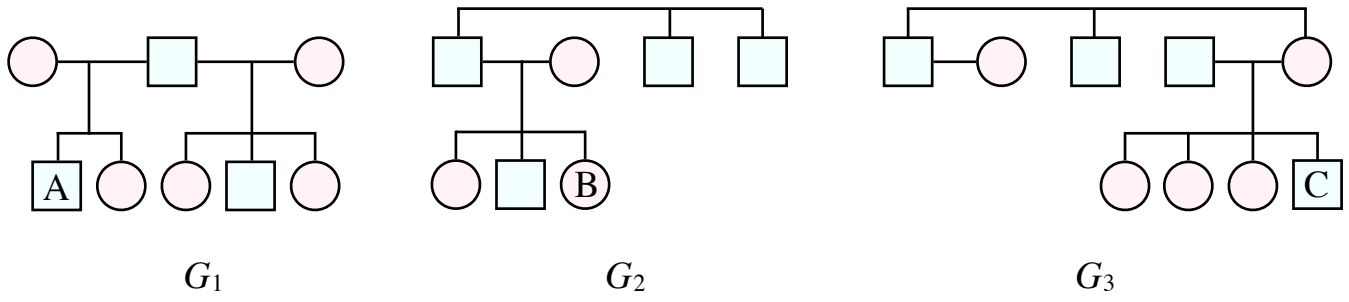
Natalie

- Theorem [Klau 2009]: The relaxed problem can be reduced to the bipartite matching problem
- for any λ, $Z_{LD}(\lambda)$ is an upper bound to the optimal alignment
  - $Z_{LD}(\lambda)$ is a relaxation
- x part of $Z_{LD}(\lambda)$ solution is a feasible alignment
  - such a feasible alignment gives a lower bound
- thus, solving $Z_{LD}(\lambda)$ gives an upper and a lower bound
- solve $\min_\lambda Z_{LD}(\lambda)$ to find the tightest upper bound
- Natalie does it with *subgradient optimization* and *dual descent*

multiple network alignment

- geneology researchers construct their own family trees
- merging these trees gives a more complete family tree, better access to information, and possibly eliminates errors

example



multiple network alignment

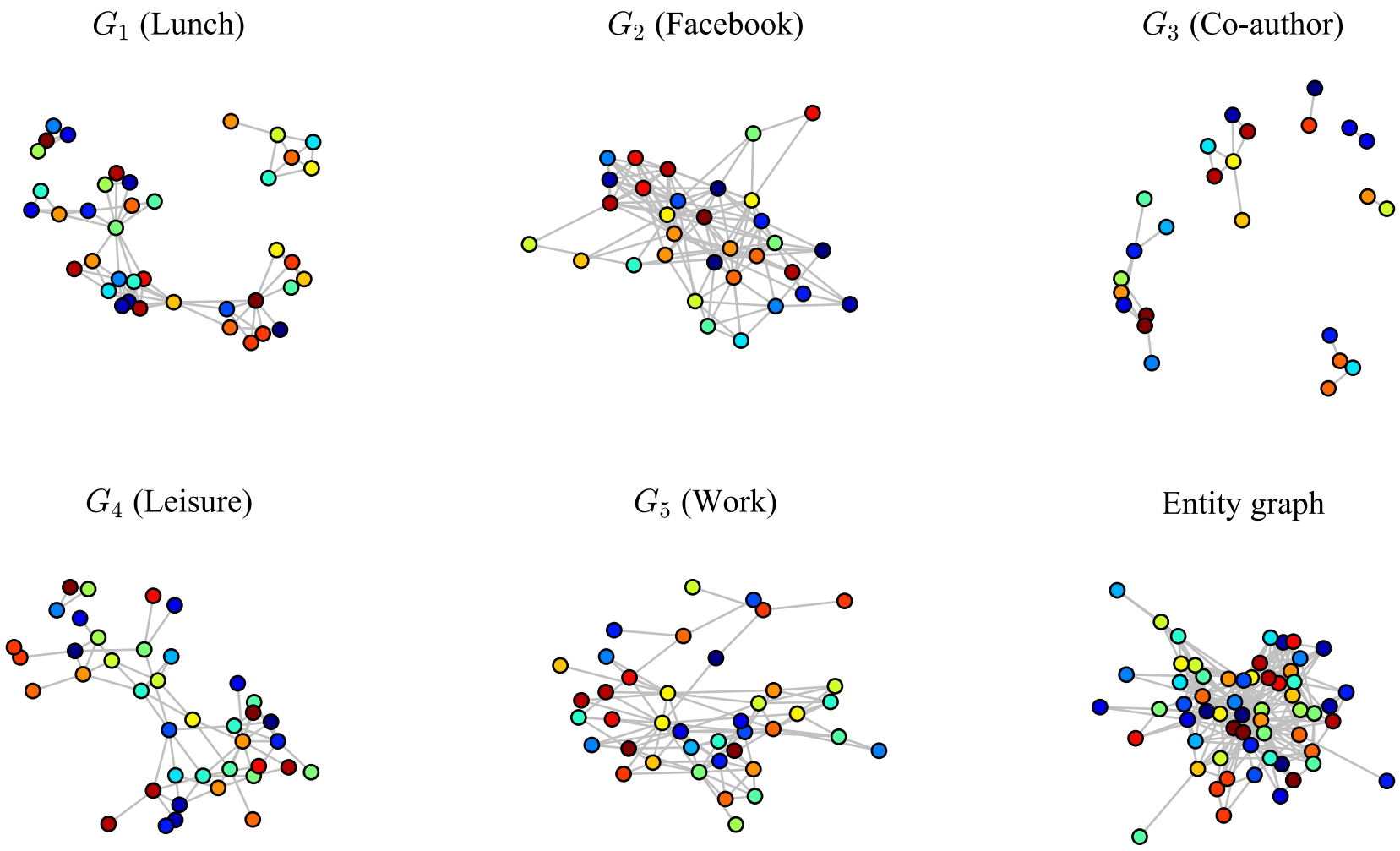


Fig. 4 An instance of the social network alignment problem. Vertices can only be aligned to other vertices with the same color. Entity graph depicts the set of underlying entities and edges between them.

multiple network alignment

- how to formalize the network-alignment problem for multiple networks?
- considered extensions of Natalie, including a formulation based on facility location

"Lagrangian relaxations for multiple network alignment"
E. Malmi, S. Chawla, A. Gionis, DMKD 2017

active network alignment

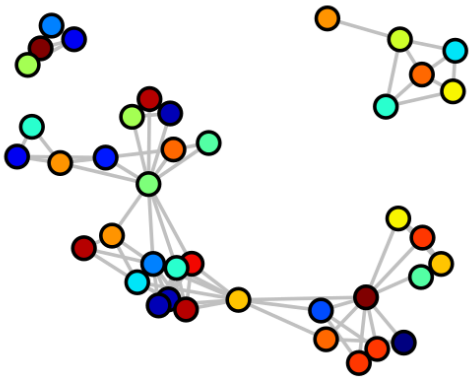
- consider : human experts can assist in network alignment
- how to optimally leverage human expertise?
- goal : ask the most informative queries from human experts

"Active network alignment: a matching-based approach"
E. Malmi, E. Terzi, A. Gionis, CIKM 2017

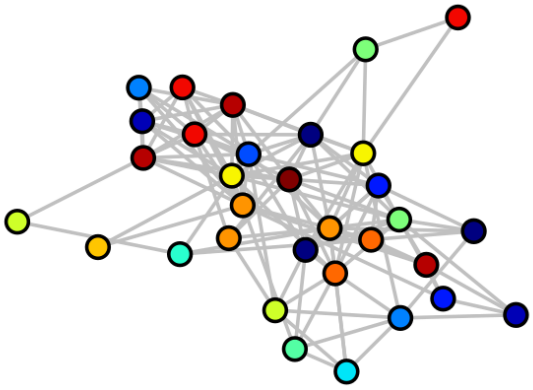
example

- what questions to ask from the human?

G_1 (Lunch)



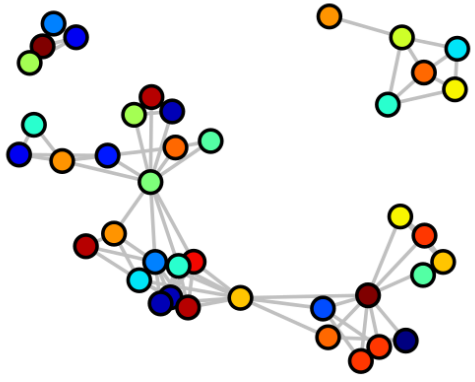
G_2 (Facebook)



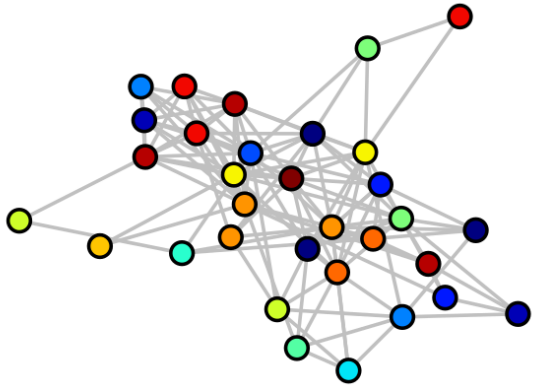
example

- what questions to ask from the human?

G_1 (Lunch)



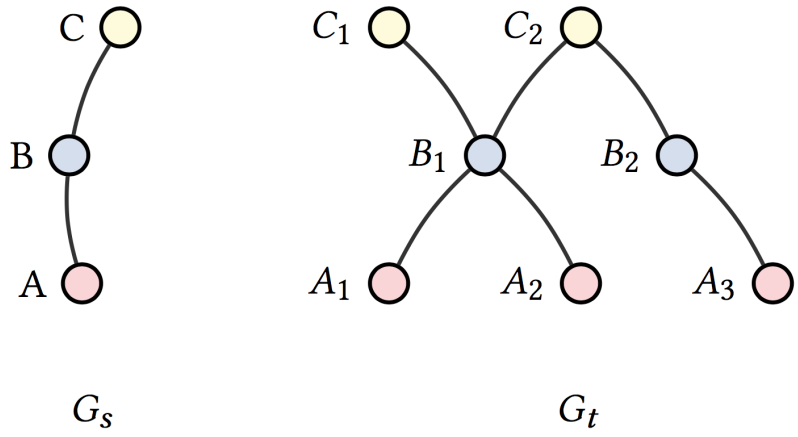
G_2 (Facebook)



- several possible query strategies
  - e.g., *"are nodes a and b the same or not?"*
  - in our work : *"which node in the set* $\{b_1, \ldots, b_n\}$ *is the most similar to node a?"*
- which nodes? the most uncertain, the most central,...

example

should an algorithm query node A, B, or C?



idea

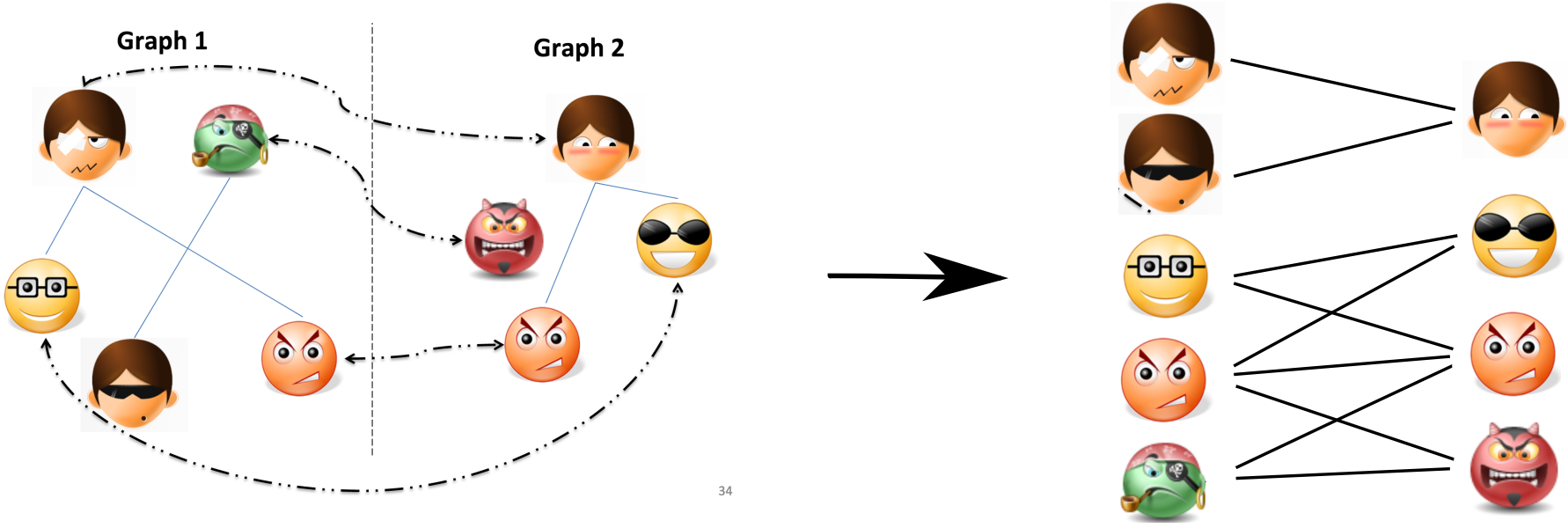
- query the most uncertain node
- how to formalize the idea?

idea

- query the most uncertain node
- how to formalize the idea?

1. consider network aligment expressed as bipartite matching (e.g., Natalie)
2. obtain a set of near-optimal matchings
3. compute the distribution of matches of a given node
4. quantify uncertainty using this distribution

step 1: network alignment as bipartite matching



step 2: sampling matchings

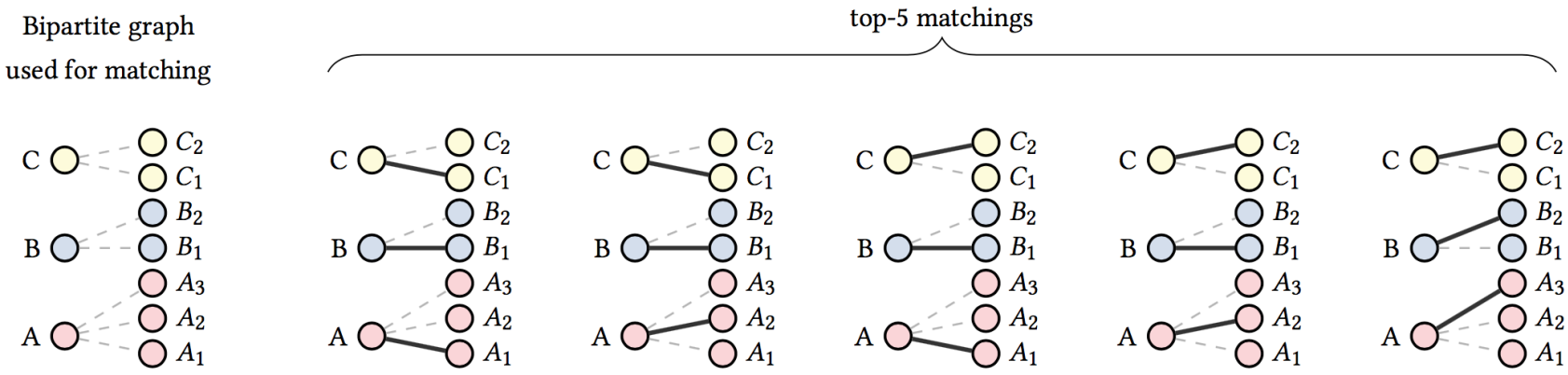
approach 1

- sample ℓ matching
- matching M sampled with probability prop. to $\exp(-\frac{1}{\beta}E(M))$
- apply a Gibbs sampler [Volkovs and Zemel, 2012]

approach 2

- find top-ℓ matchings
- apply Murthy's algorithm [Murthy, 1968]
- running time $O(\ell n^3)$ [Miller et al., 1997]

steps 1–3: example



The matching distributions for A, B, and C are

$\{A_1 : 40\%, A_2 : 40\%, A_3 : 20\%\}$,

$\{B_1 : 80\%, B_2 : 20\%\}$, and

$\{C_1 : 40\%, C_2 : 60\%\}$, respectively.

quantifying uncertainty

- let $f_v(u)$ denote how many times node $v \in V_1$ has been assigned to $u \in V_2$ among the sampled matchings

  approach 1

- select to query node $\hat{v} = \arg\min_v \max_u f_v(u)$

  approach 2

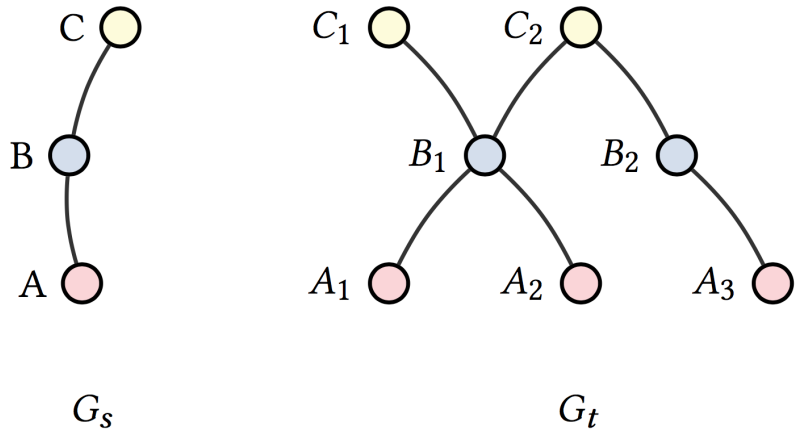
- define uncertainty using entropy

  approach 3

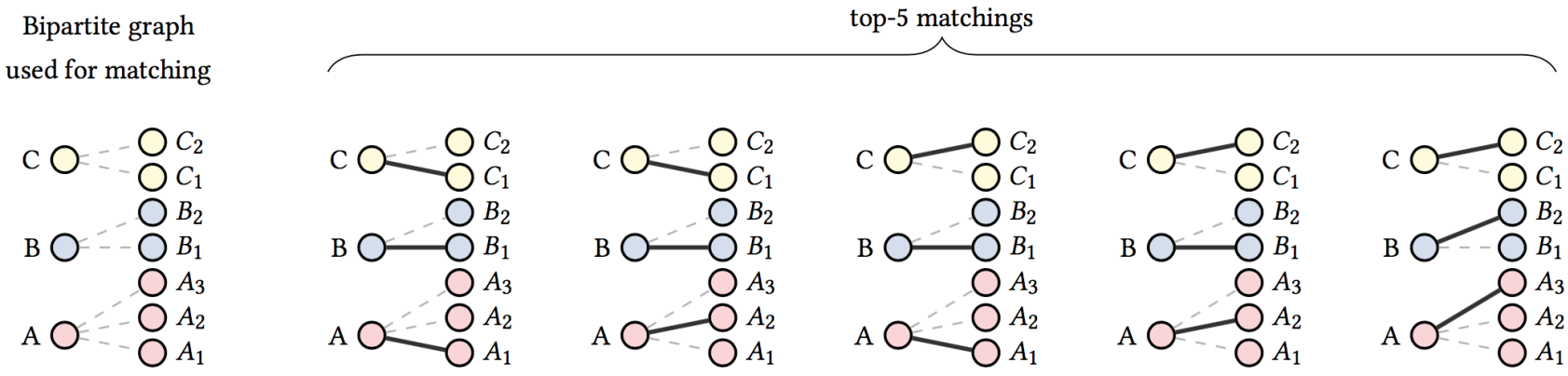
- select to query node $\hat{v}$ that maximizes the expected certainty of the remaining nodes

example

should an algorithm query node A, B, or C?



quantifying uncertainty : example



- query node $\hat{v} = \arg\min_v \max_u f_v(u)$

  $\{A_1 : 40\%, A_2 : 40\%, A_3 : 20\%\} \rightarrow$ query node A

  $\{B_1 : 80\%, B_2 : 20\%\}$

  $\{C_1 : 40\%, C_2 : 60\%\}$

experiments

datasets

- preferential-attachment graphs
- social networks
- genealogical networks

experiments

datasets

- preferential-attachment graphs
- social networks
- genealogical networks

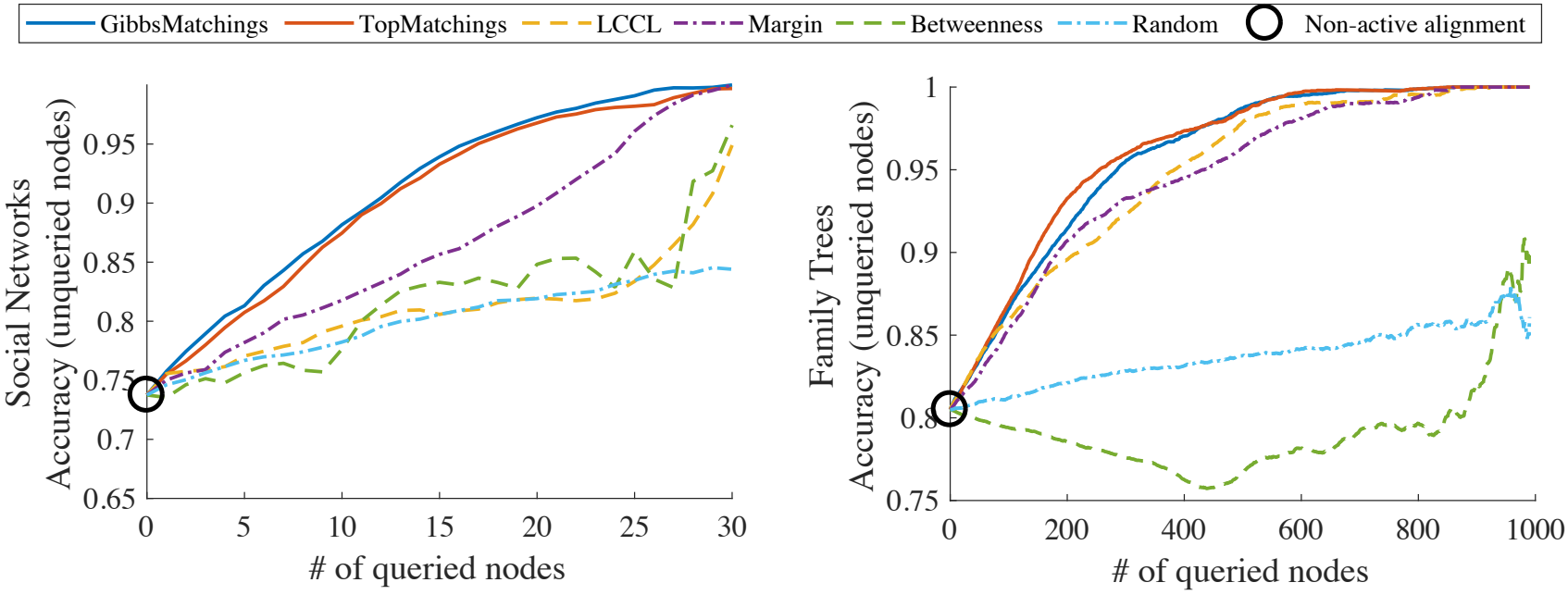
evaluation

1. run a non-active aligner
2. compute accuracy (% of correctly aligned non-queried nodes)
3. query a node
4. go back to step 1

baselines

- **random**: query random node
- **betweenness**: largest betweenness centrality
- **margin**: smallest difference between the top-2 nodes
- **LCCL**: least confident given the current labelling
  [Cortés & Serratosa 2013]

results



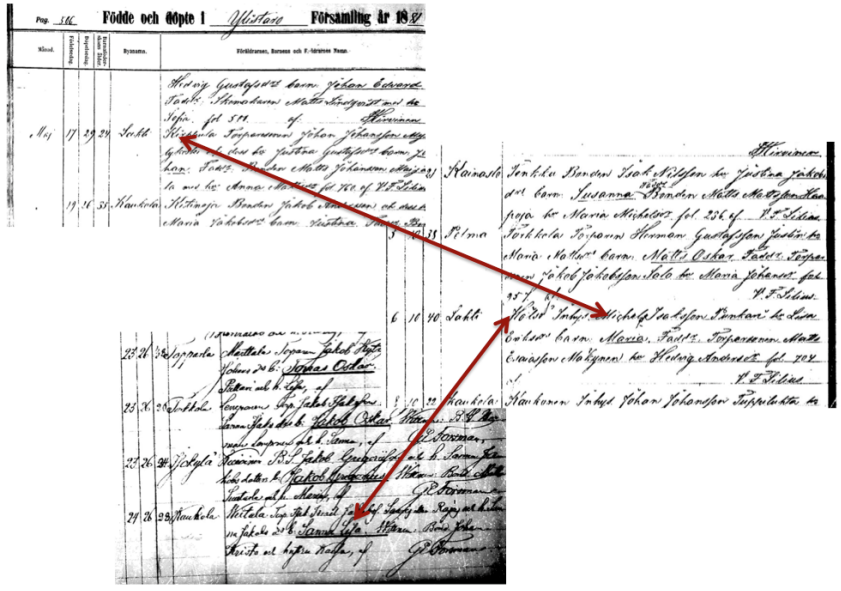
the proposed methods, TopMatchings and GibbsMatchings, outperform the baseline query strategies. Accuracy is the % of correctly aligned unqueried nodes using Natalie [Klau 2009]

conclusions

- active network alignment to leverage human experts
- main novelty: the combination of
  (1) viewing network alignment as bipartite matching +
  (2) sampling matchings +
  (3) quantifying uncertainty based on marginal distributions
- applicable on top of any matching-based aligner
- future work
  (1) relative vs. absolute queries
  (2) imperfect oracle

genealogical network inference

- dataset: 10 million transcribed vital/parish records from Finland
  - births, marriages, burials, migration
  - from 1648 to 1917
- task: link birth records to the parents' birth records
- challenges:
  - duplicate names
  - spelling variations
  - missing records



linking birth records

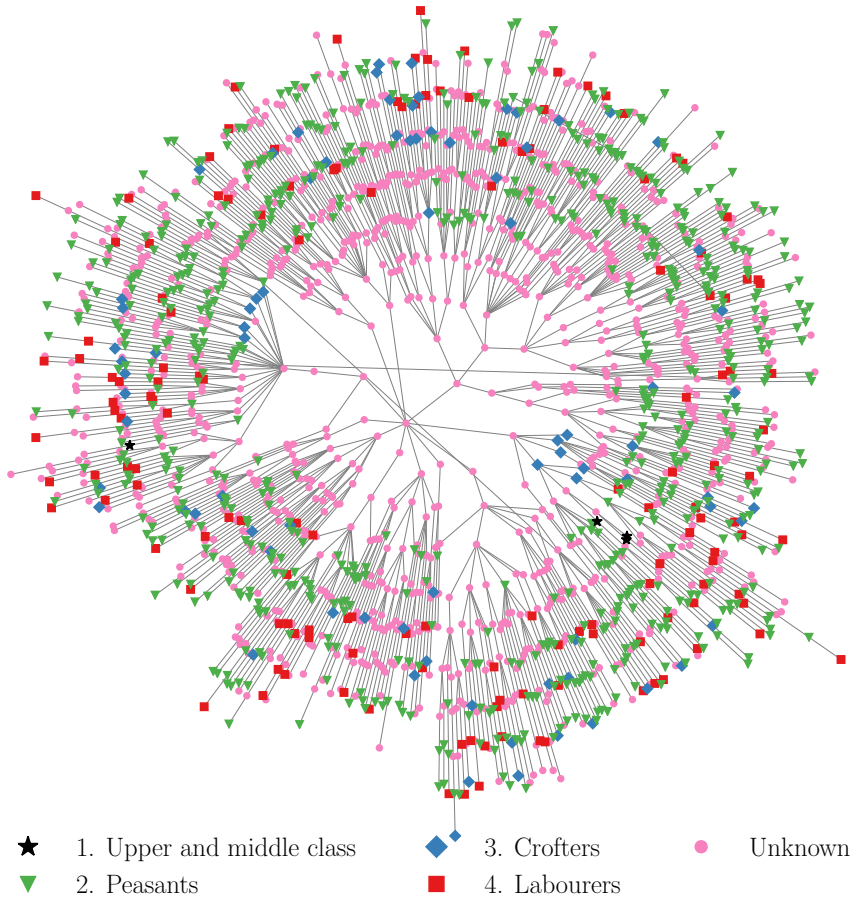
- birth record attributes: child name, parent names, birth date, birth location, parent occupations
- training/test data: 18 731 ground truth links
- approach
  1. retrieve candidate parents based on names and birth years
  2. probabilistically classify each child–parent edge to *match* vs. *no match*
  3. compute a probability distribution over the candidates

E. Malmi, M. Rasa, A. Gionis *"AncestryAI: A tool for exploring computationally inferred family trees"*, WWW Companion, 2017.

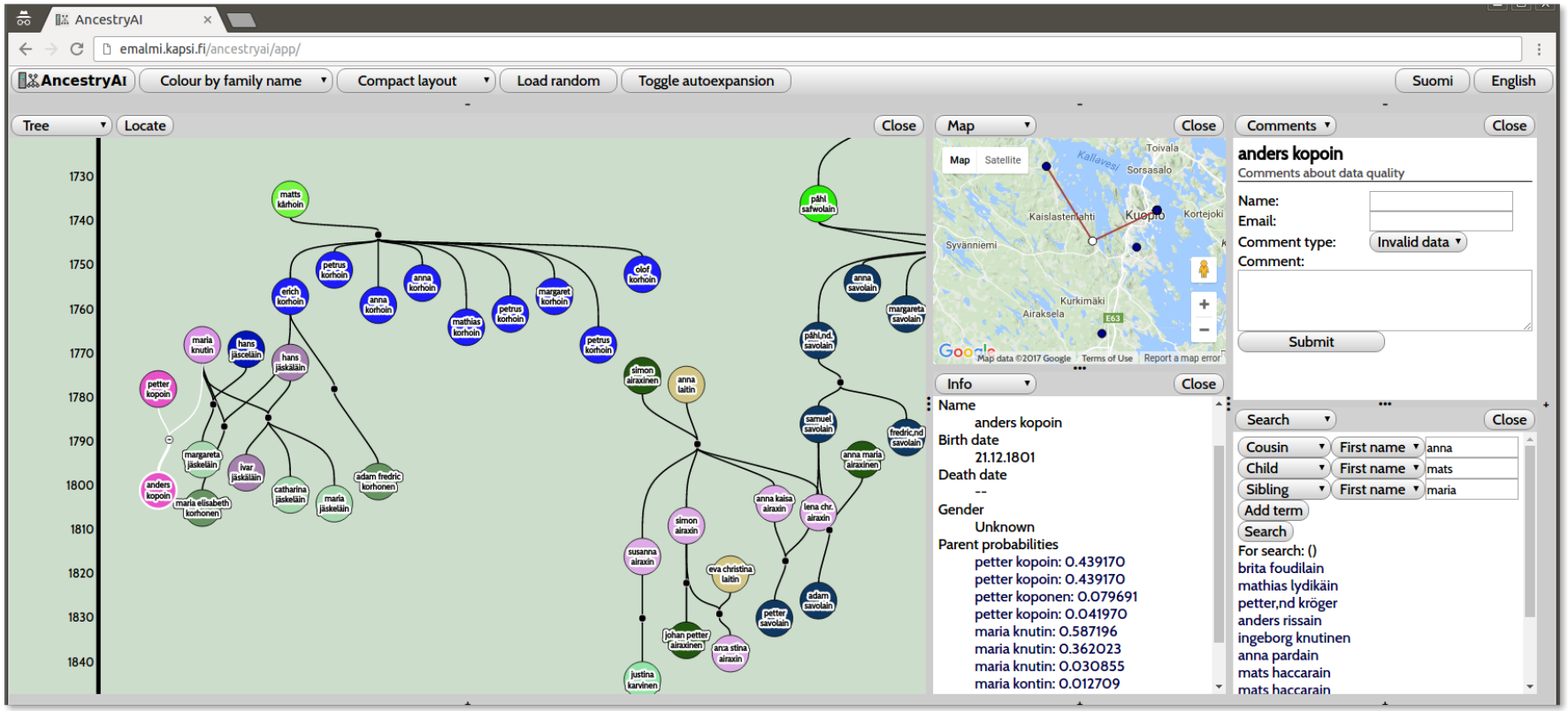
E. Malmi, A. Solin, A. Gionis *"The blind leading the blind: Network-based location estimation under uncertainty"*, ECML PKDD, 2015.

results

- largest component:
  2.6 million individuals
- subgraph of 2 000 nodes
  and 13 generations shown
  on the right



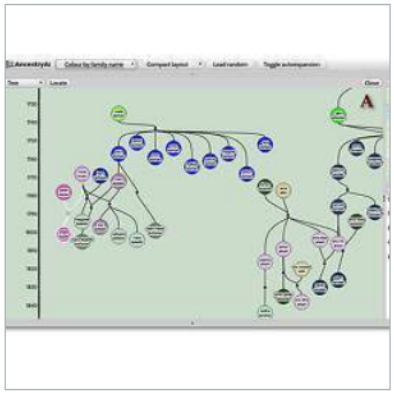
online tool: ancestryai.cs.hut.fi



ACM TECHNEWS

AncestryAI Algorithm Traces Your Family Tree Back More Than 300 Years

By Aalto University
March 17, 2017
Comments



Researchers at Finland's Aalto University have developed a family tree artificial intelligence algorithm.

Credit: Eric Malmi

Aalto University's Eric Malmi.

Researchers at Aalto University in Finland have developed AncestryAI, a family tree artificial intelligence (AI) algorithm that looks for connections between 5 million baptisms from the end of the 17th century to the beginning of the 20th century.

AncestryAI, which is part of the HisKi project, collects parish data on baptisms, marriages, and relocations. It then automatically searches for a child's most probable parents and creates family trees based on this information. The algorithm offers several options based on the parents' date, place of birth, and similar names.

In addition, AncestryAI users can leave comments on the system's accuracy, and the algorithm uses these comments to improve its analysis.

"It would be really interesting to have a family tree covering the whole of Finland, because it could also be used to study wars, epidemics, the influence of and changes in the class society," says

From *Aalto University*
View Full Article

MORE NEWS & OPINIONS

Why the Computing Cloud Will Keep Growing and Growing
The New York Times

Introducing 'Operator 4.0,' a tech-augmented human worker
The Conversation

Just Press Reboot
Bertrand Meyer

case study : assortative mating

- assortative mating

  = marrying your like

- instance of social stratification

case study : assortative mating

- assortative mating

  = marrying your like

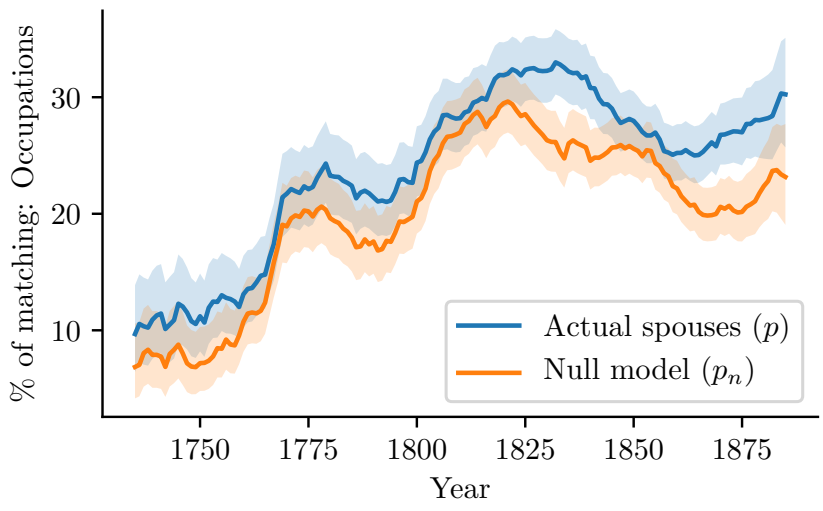
- instance of social stratification
- research questions:

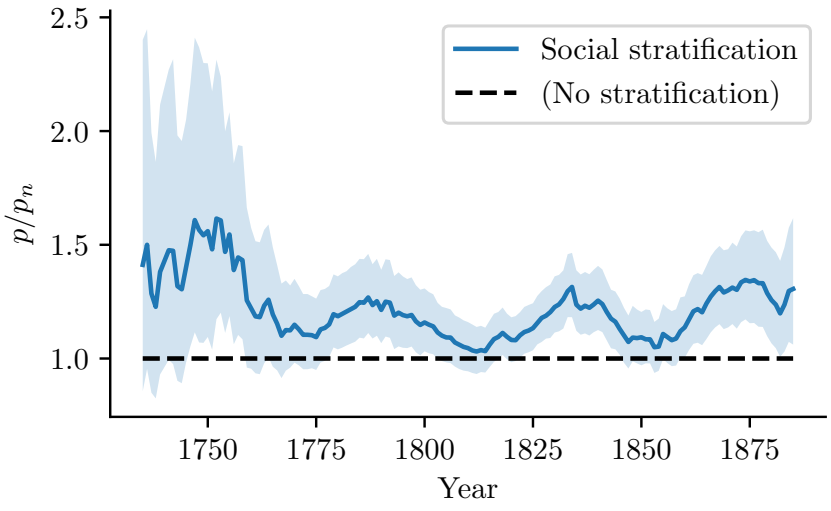
  (*i*) can we detect assortative mating in the inferred genealogical network?

  (*ii*) how does it change over time?

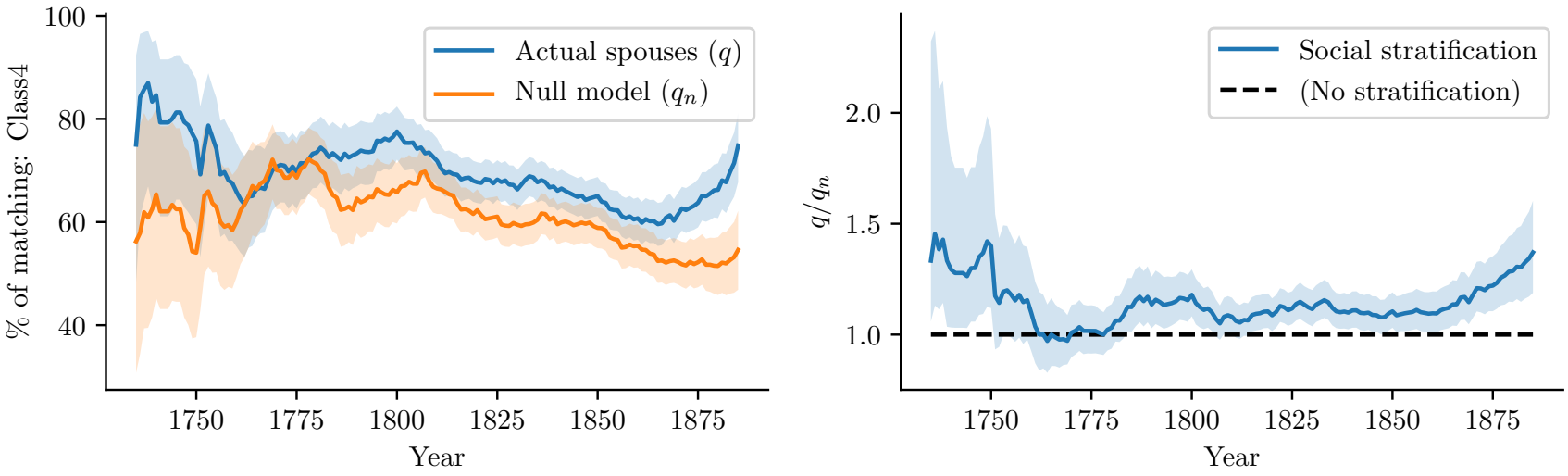
assortative mating results

- probability of matching spouse father occupations
- null model: shuffle spouses





assortative mating results (occupations clustered)



Class4: (1) upper and middle class, (2) peasants, (3) crofters, and (4) labourers

conclusions

1. accurate links
2. link probabilities
3. assortative mating
   - (*i*) did occur in Finland (1735–1885)
   - (*ii*) did not monotonously decrease or increase
4. longitudinal computational social science

thank you!

credits

Eric
Malmi

Evimaria
Terzi

Sanjay
Chawla