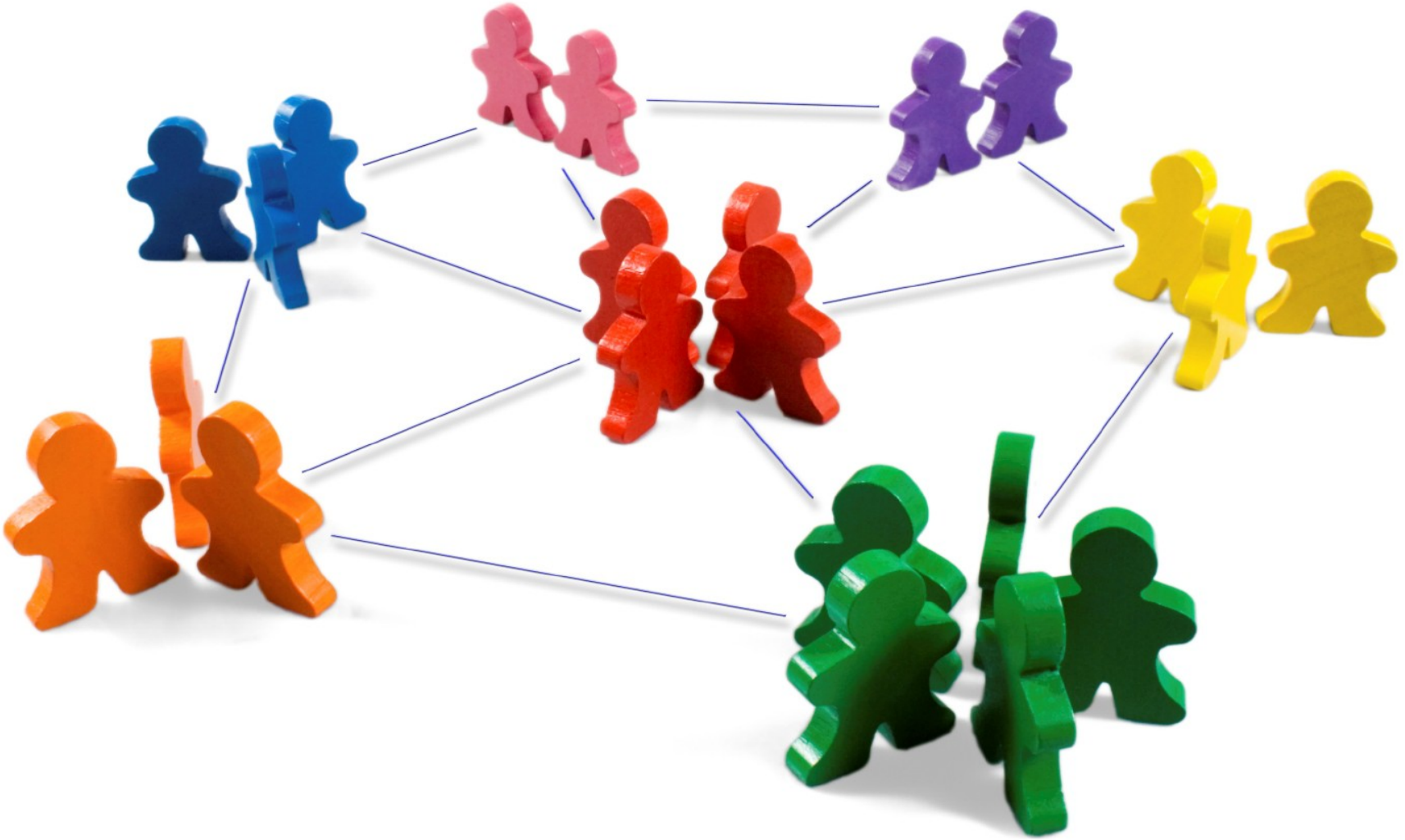
Cross Document Coreference Resolution with Diffusion based Community Detection

Kambiz Ghoorchian
Sarunas Girdzijauskas
Royal Institute of Technology – KTH
iSocial

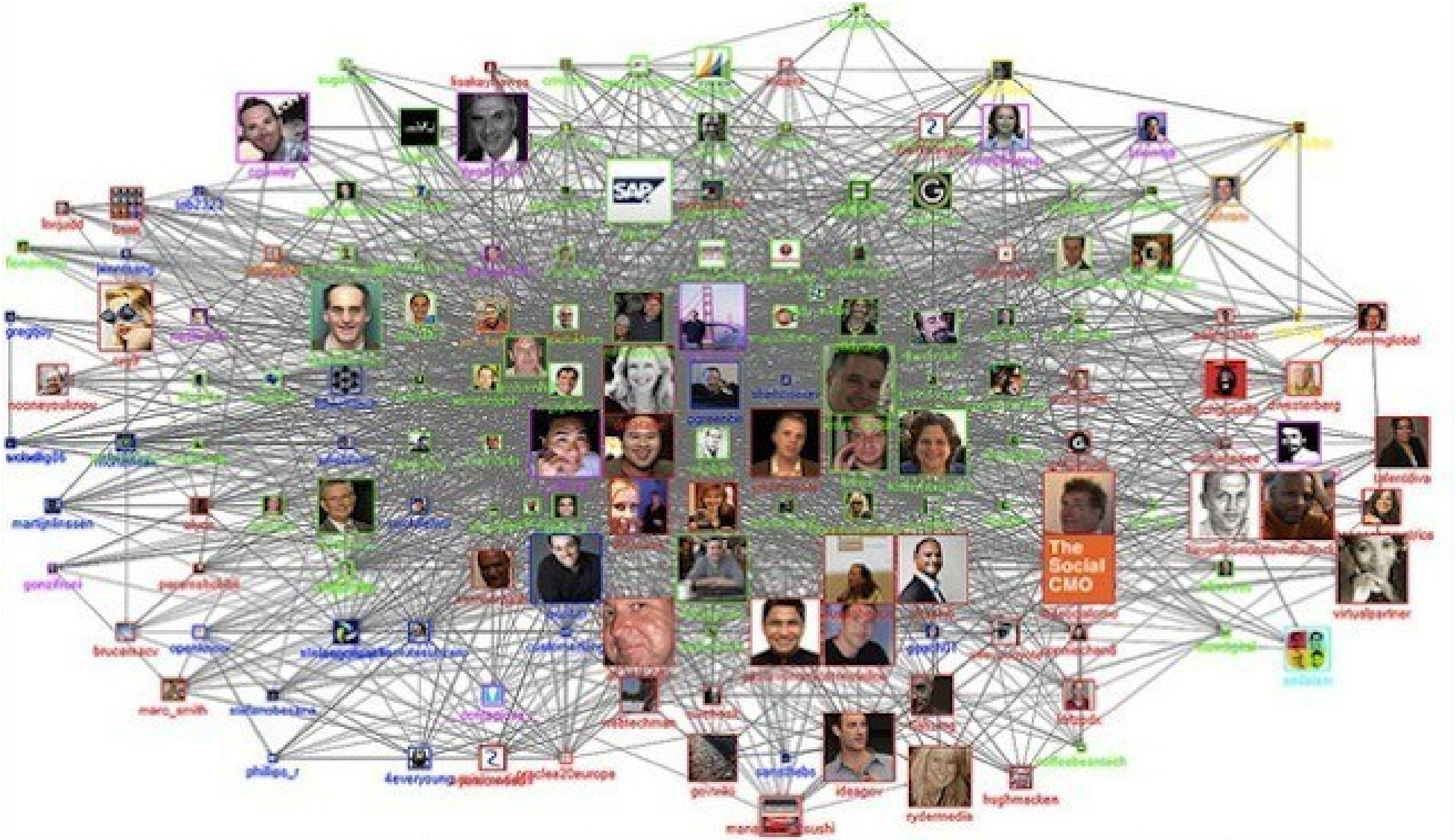
February – 2014

kambiz.ghoorchian@gmail.com

Community Detection



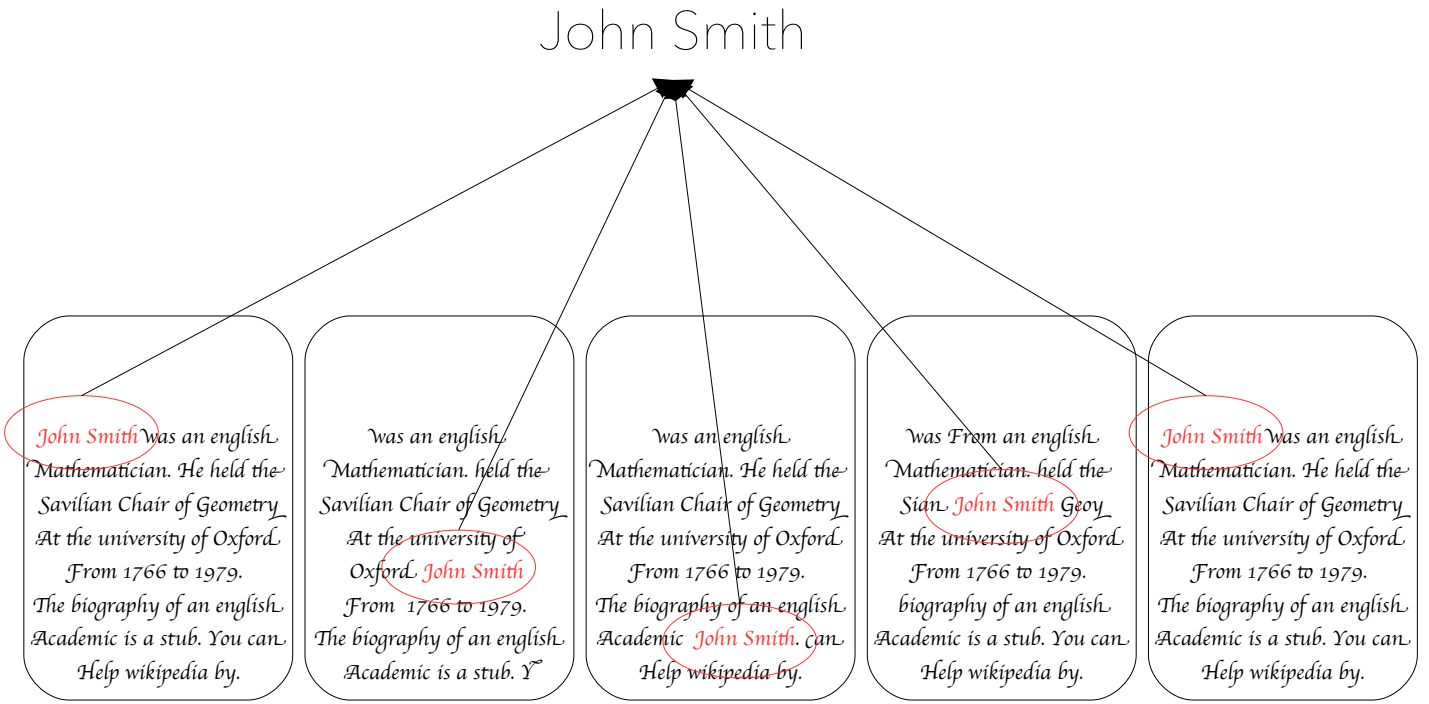
Community Detection



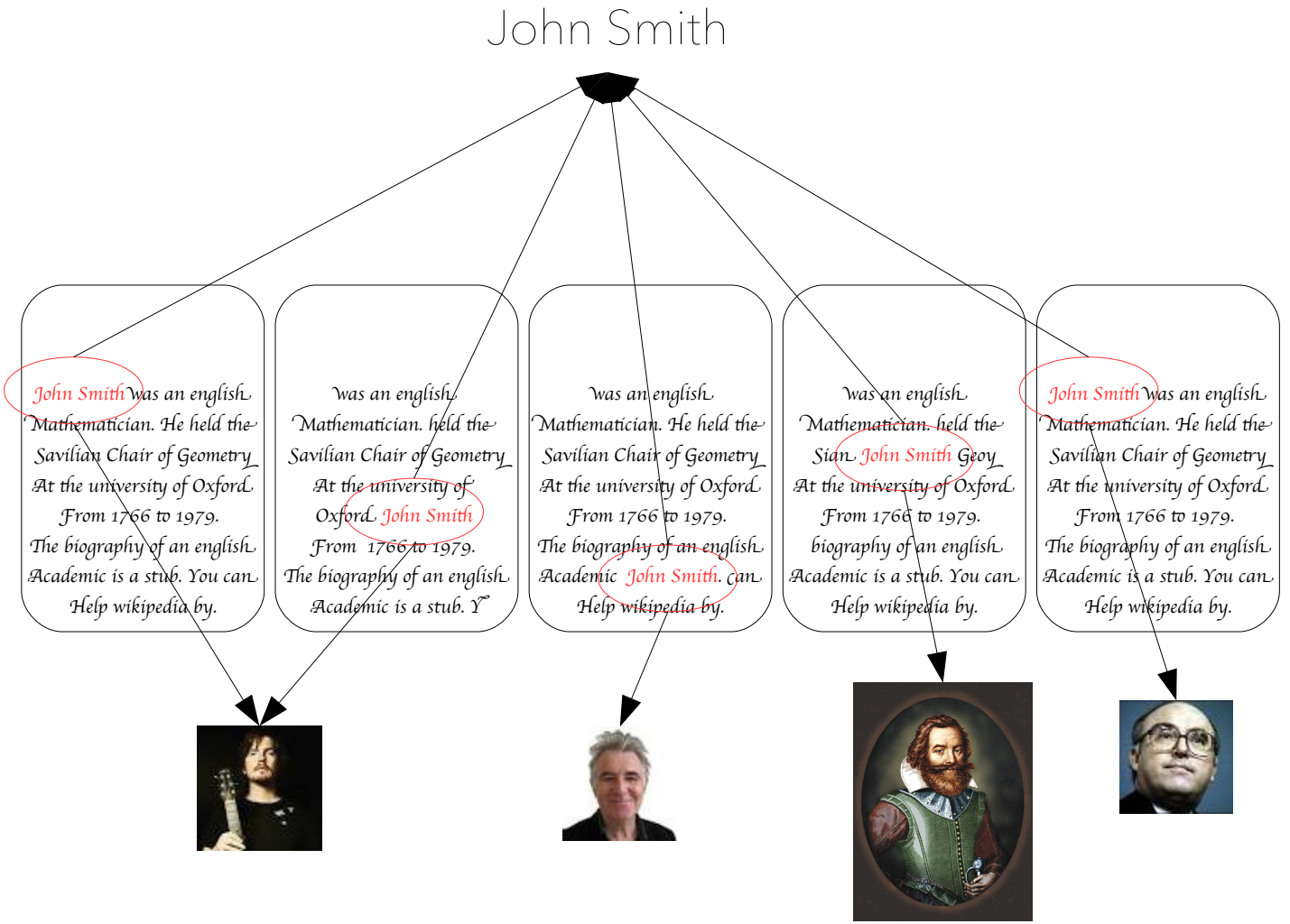
Overview

- Coreference Resolution
- Previous Solutions
- Limitations
- Community Detection
- Model
- Algorithm
- Results & Conclusion

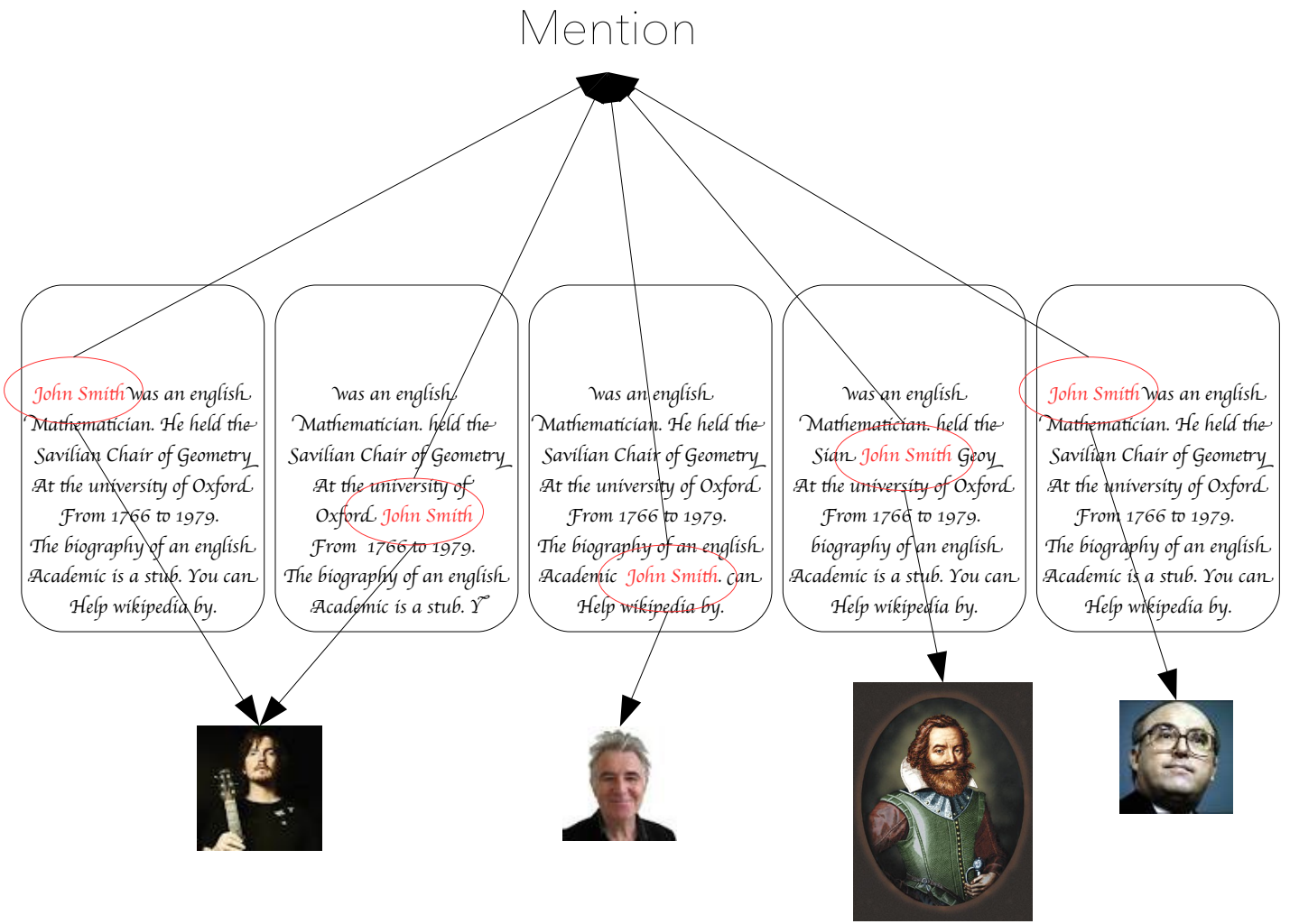
Coreference



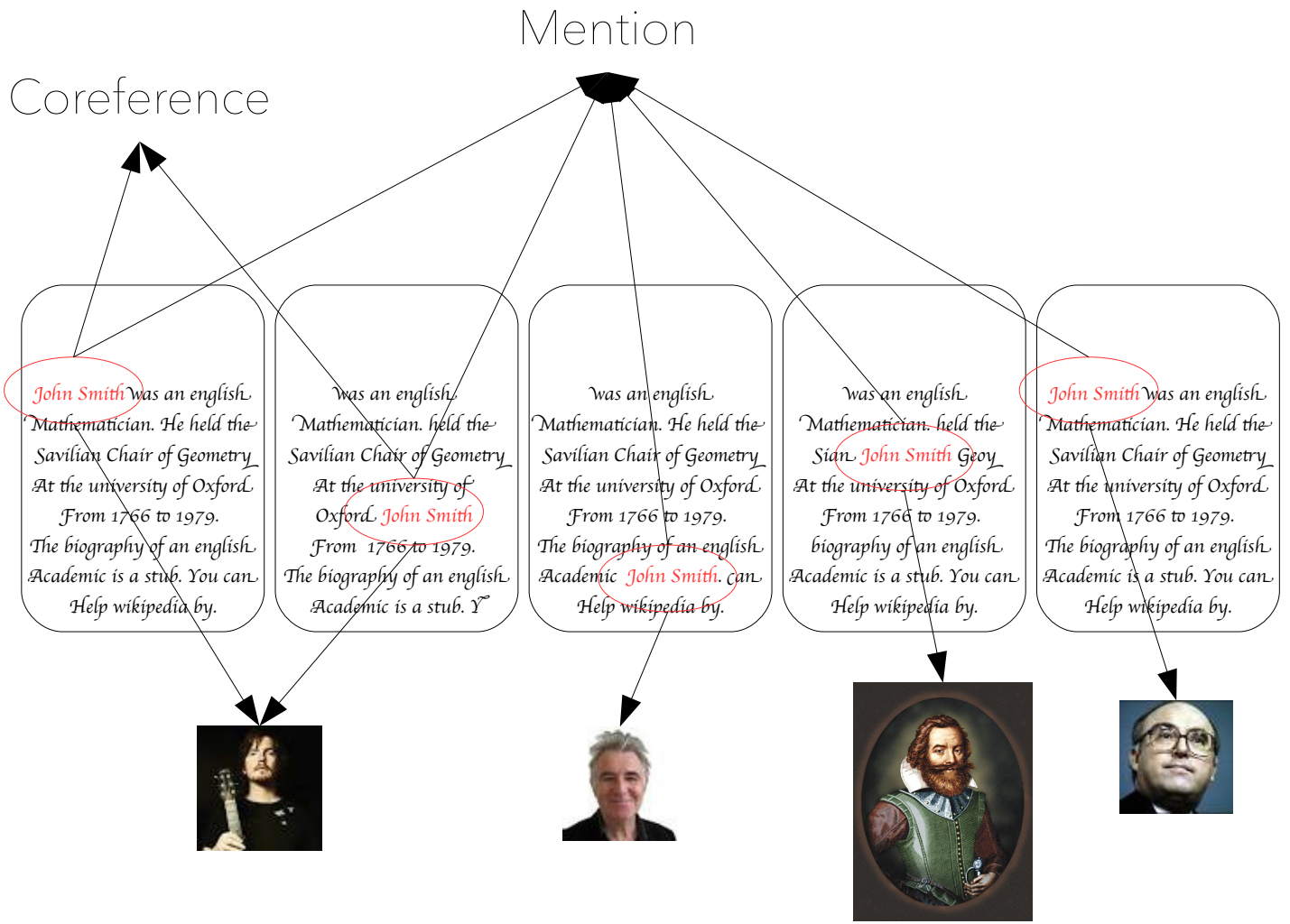
Coreference



Coreference



Coreference

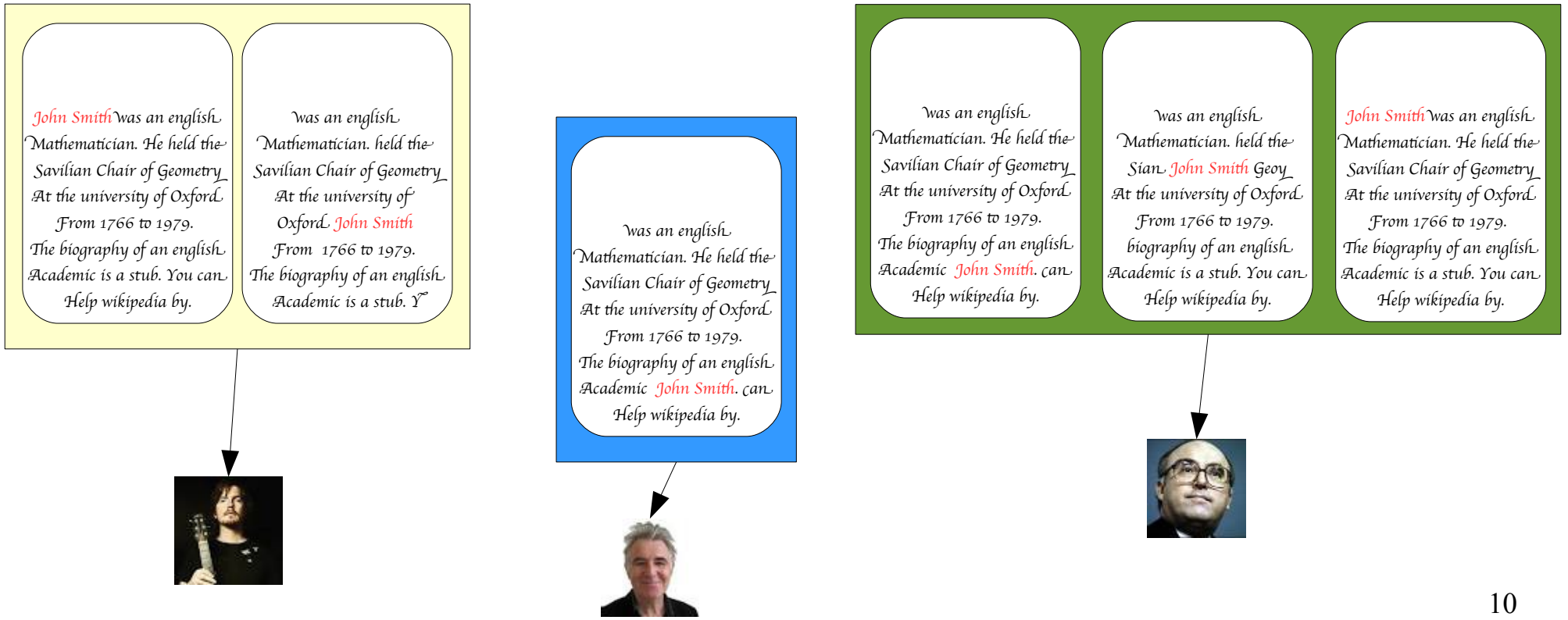


Definitions

- Entity
  - A Distinctive and Independent thing in real world.
- Mention
  - A linguistic phenomena (Word or Phrase) referring to an entity.
- Coreference:
  - Two or more mentions referring to the same entity.

Coreference Resolution

- Is the task of clustering multiple documents based on the Entity they are talking about.



Greedy Clustering

- Vector space model

	W1	W2	W3	...	Wn
D1	1	1	0	...	1
D2	1	0	1	...	0
D3	1	1	1	...	0
...	...	...	...	...	...
Dn	0	0	1	...	1

- Pair wise Mention similarity
  - Cosine, Jacard Index, Hamming Distance, TFIDF
- Heuristic Feature sets
  - Text, Context or Document level feature vectors

Greedy Clustering

- Pair wise mention similarity
  - O(N^2)

	Documents	Mentions	Entities
John Smith	197	197	35
Person-X	34,404	34,404	14,767

Greedy Clustering

- Pair wise mention similarity
  - O(N^2)

	Documents	Mentions	Entities
John Smith	197	197	35
Person-X	34,404	34,404	14,767
wiki-link	10,839,248	40,323,863	2,933,659

Greedy Clustering

- Pair wise mention similarity
  - O(N^2)

	Documents	Mentions	Entities
John Smith	197	197	35
Person-X	34,404	34,404	14,767
wiki-link	10,839,248	40,323,863	2,933,659

Distributed Clustering

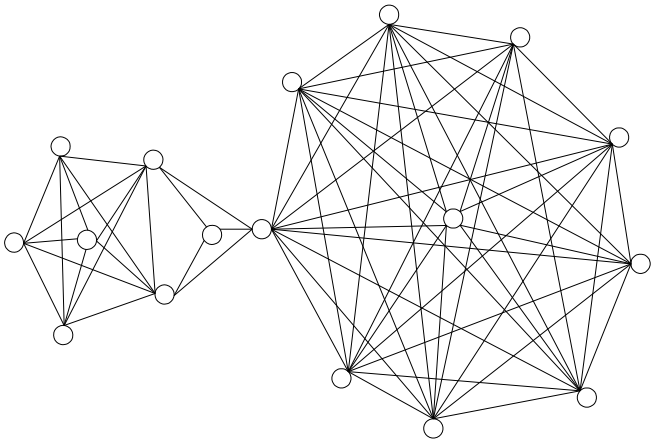
Diffusion based Community Detection

Our Solution

I. F.Rahimian, Sarunas Girdzijauskas, Seif Haridi. Distributed Community Detection for Large-Scale Coreference Resolution, 2014. waiting for the commete decision.

Diffusion based Community Detection

Graph vs Vector



	W1	W2	W3	...	Wn
D1	1	1	0	...	1
D2	1	0	1	...	0
D3	1	1	1	...	0
...	...	...	...	...	...
Dn	0	0	1	...	1

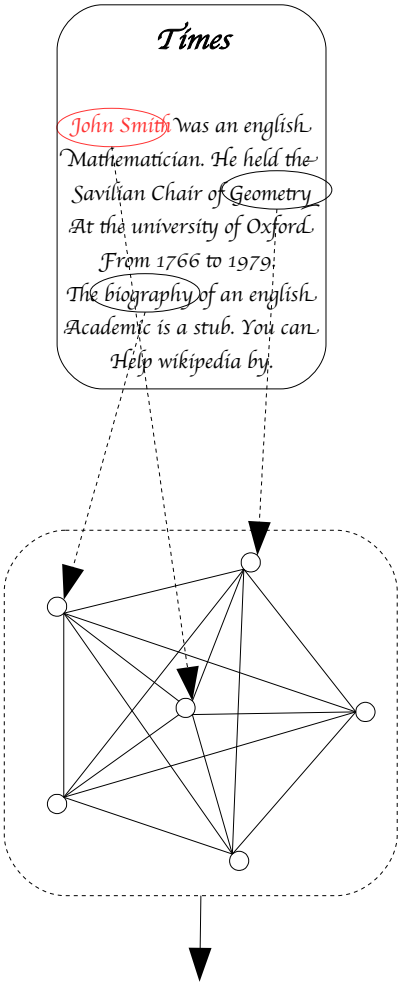
Diffusion based Community Detection

Adjacency Matrix vs Similarity Matrix

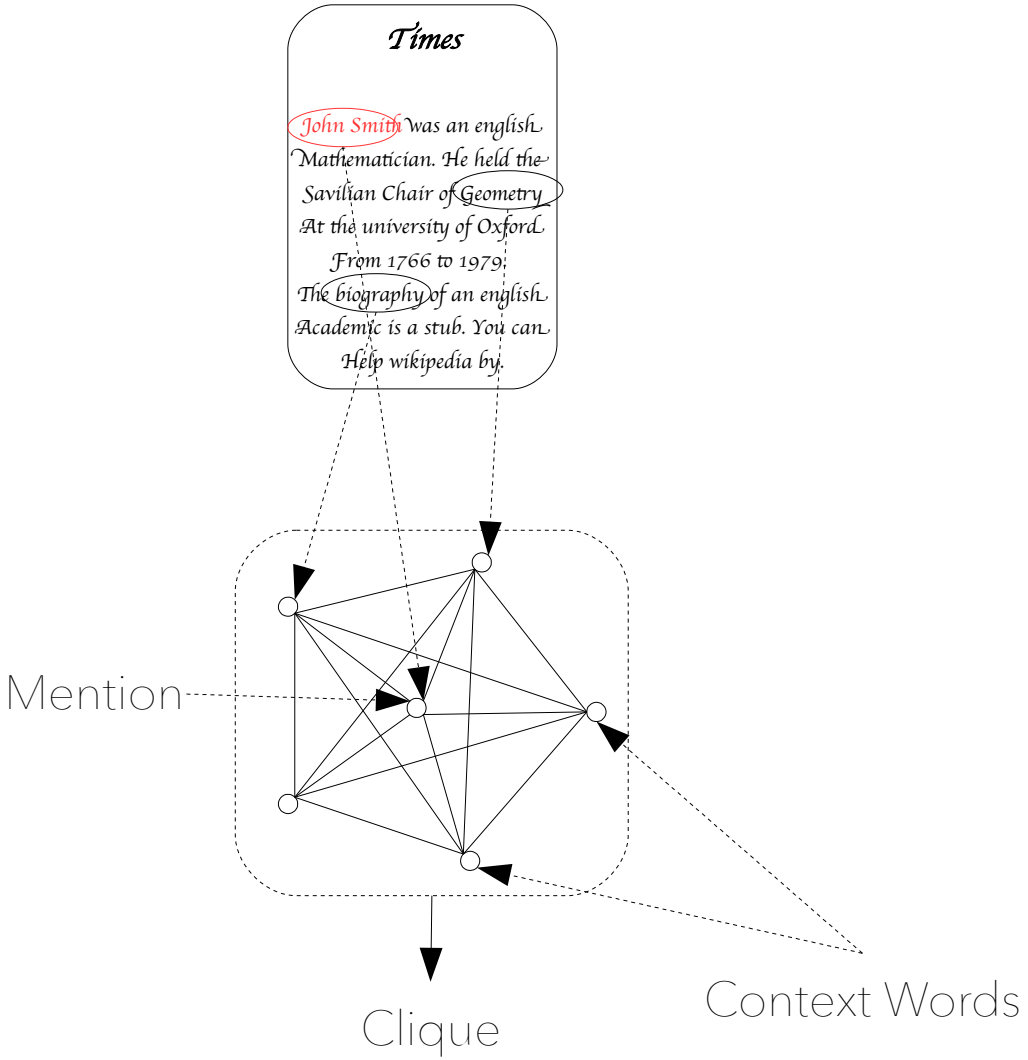
	0	**1**	**2**	**...**	**n**
0	0	1	0	...	1
1	1	0	1	...	0
2	1	1	0	...	0
...	...	...	...	...	...
n	0	0	1	...	0

	0	**1**	**2**	**...**	**n**
0	0	2	6	...	m1
1	2	0	1	...	m2
2	6	1	0	...	m3
...	...	...	...	...	...
n	m1	m2	m3	...	0

Model


Times
John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.
Clique

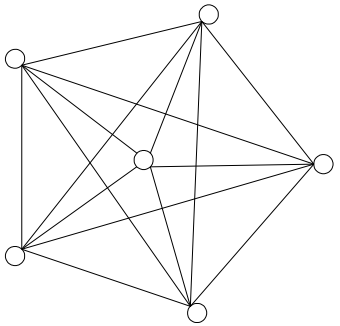
Model


Times
John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.
Mention
Clique
Context Words

Model

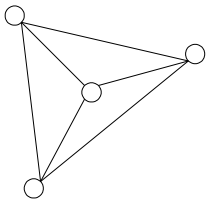
Times

John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



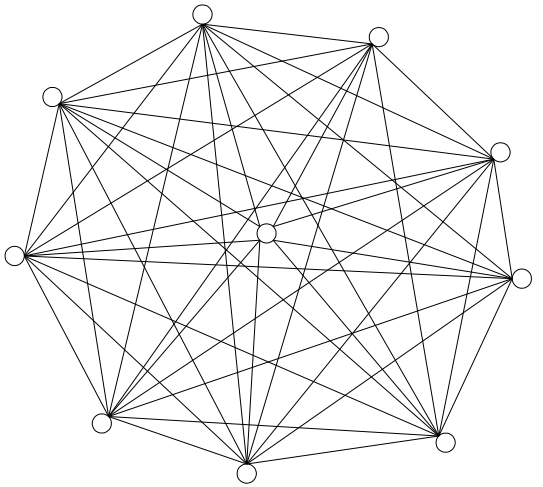
Times

was an english Mathematician. He held the Savilian John Smith Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can



Times

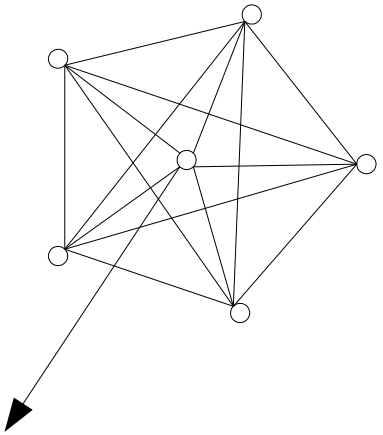
was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford John Smith to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



Model

Times

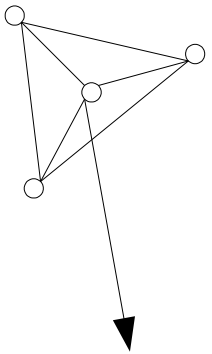
John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



JohnSmith:1

Times

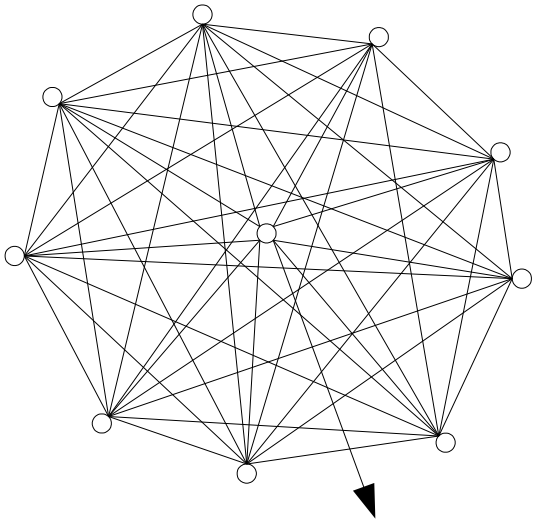
was an english Mathematician. He held the Savilian John Smith Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can



JohnSmith:2

Times

was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford John Smith to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



JohnSmith:3

Model

Times

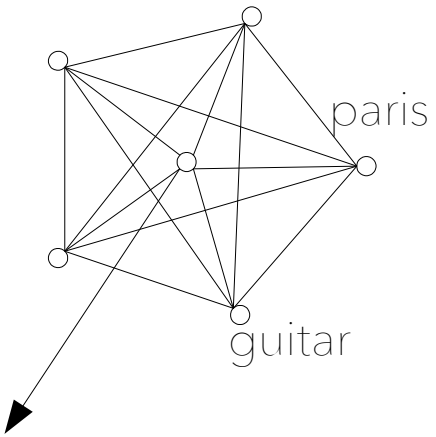
John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.

Times

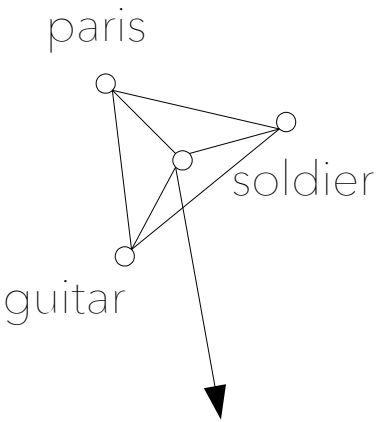
was an english Mathematician. He held the Savilian John Smith Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can

Times

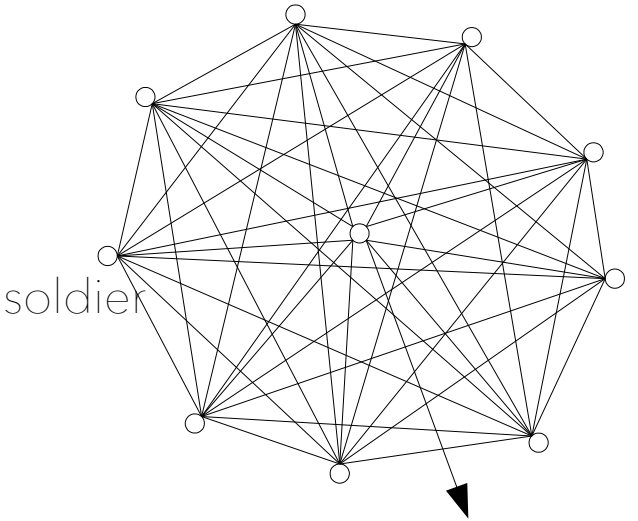
was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford John Smith to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



JohnSmith:1



JohnSmith:2



JohnSmith:3

Model

Times

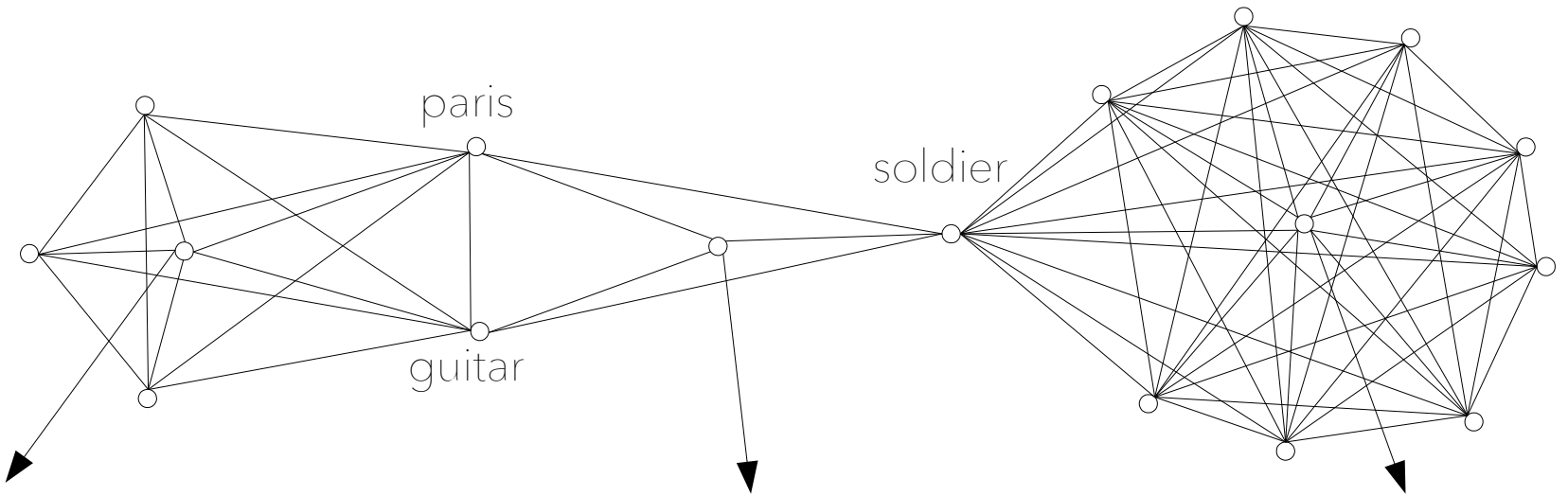
John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.

Times

was an english Mathematician. He held the Savilian John Smith Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can

Times

was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford John Smith to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.



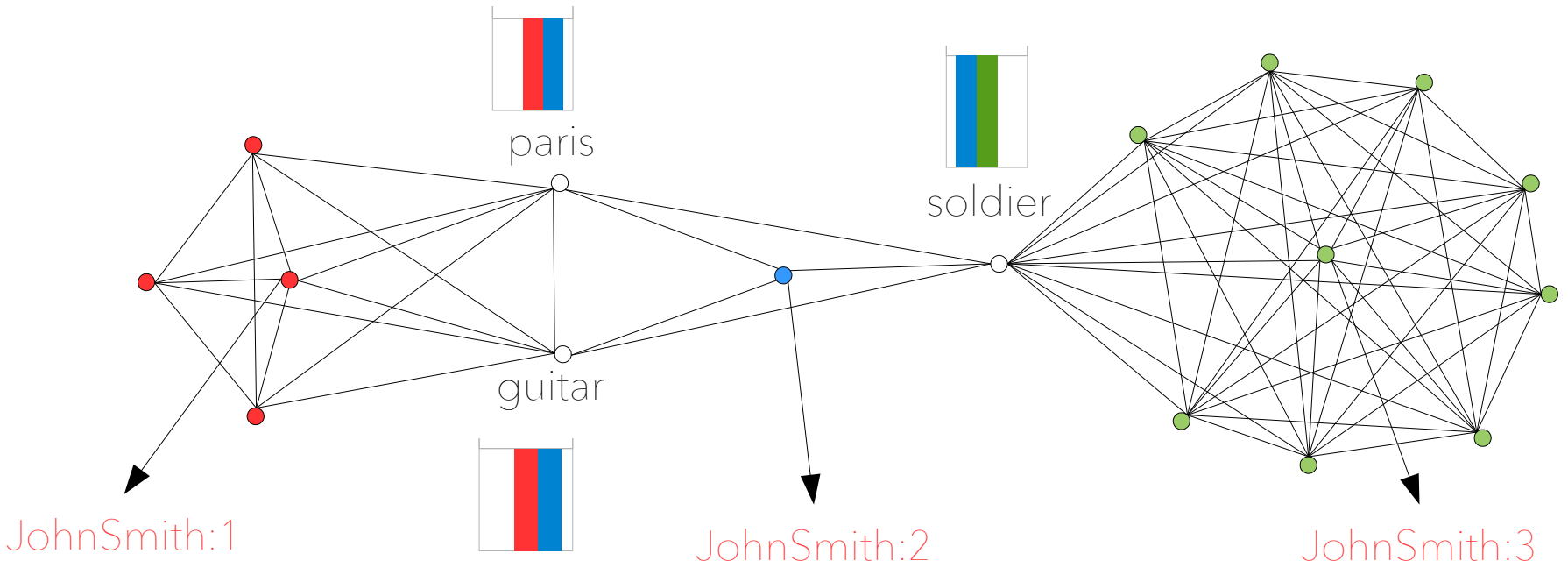
JohnSmith:1 JohnSmith:2 JohnSmith:3

Algorithm - Initialization

John Smith was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.

was an english Mathematician. He held the Savilian John Smith Chair of Geometry At the university of Oxford From 1766 to 1979. The biography of an english Academic is a stub. You can

was an english Mathematician. He held the Savilian Chair of Geometry At the university of Oxford John Smith to 1979. The biography of an english Academic is a stub. You can Help wikipedia by.

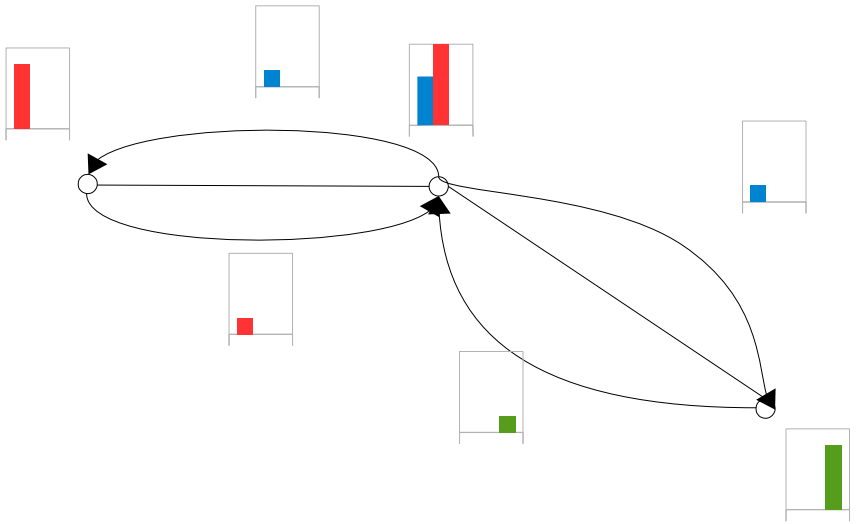


Algorithm - Iteration

- Gather
- Apply
- Scatter

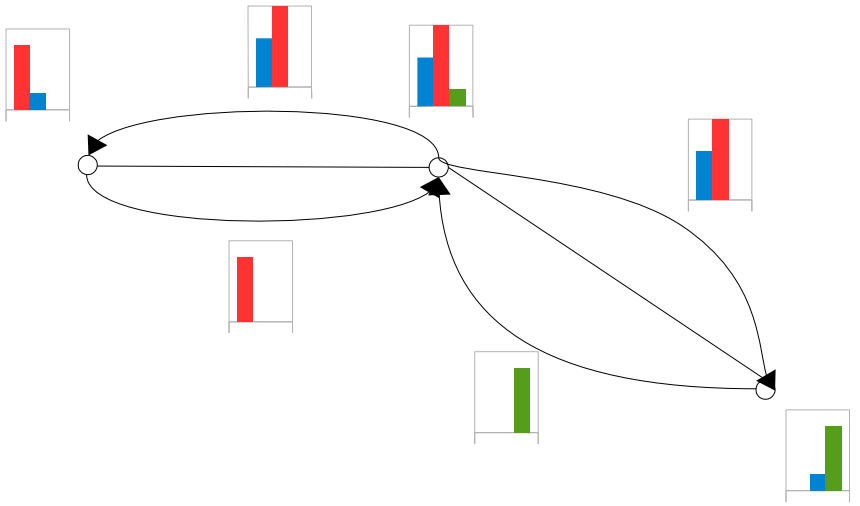
Algorithm - Iteration

- Gather
  - Share



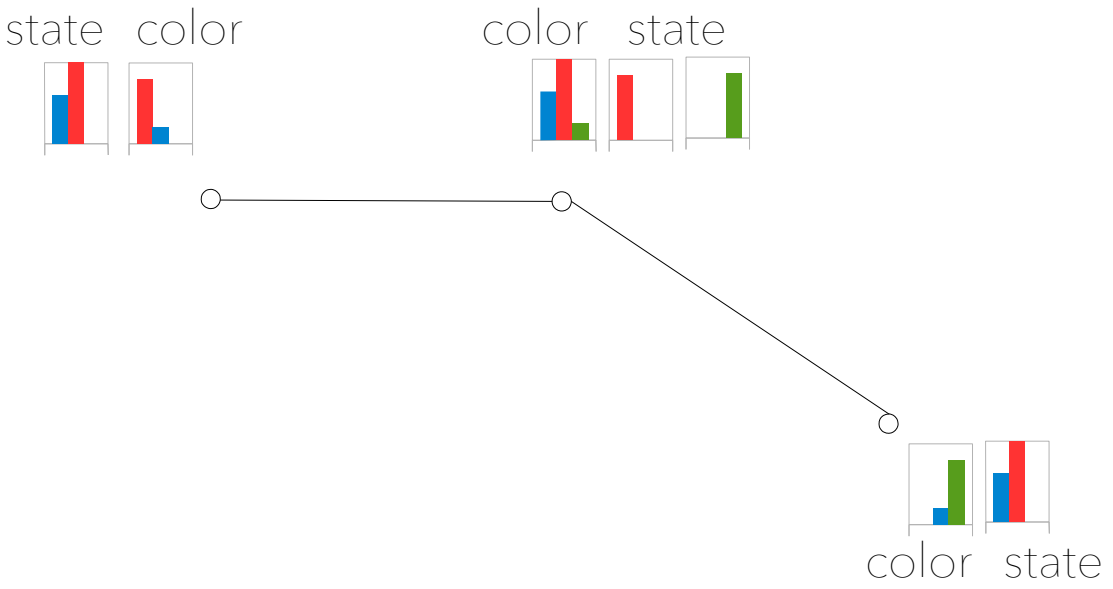
Algorithm - Iteration

- Gather
  - Share
  - Neighbor's State



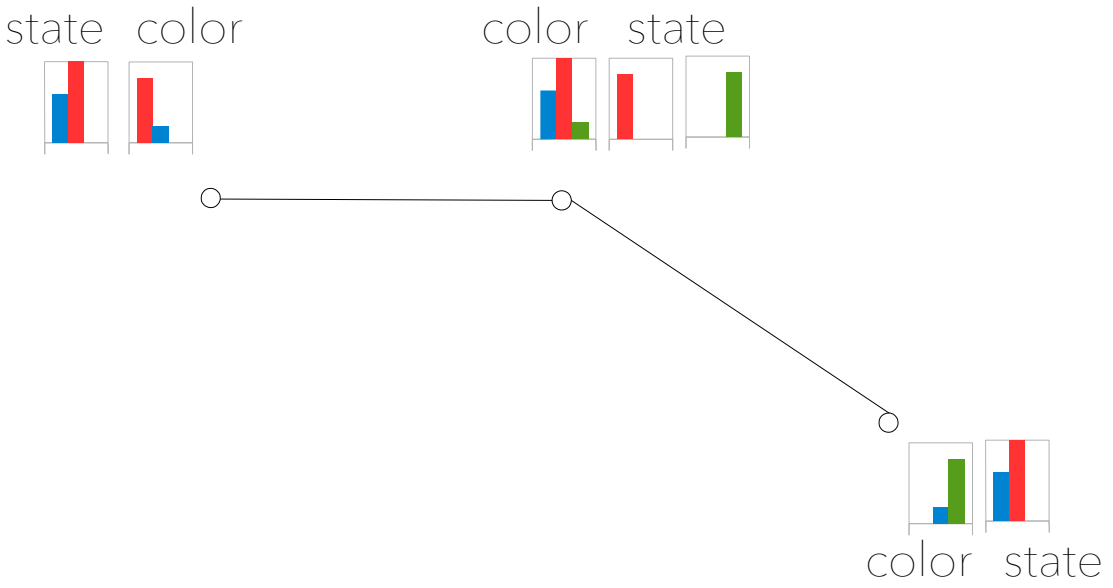
Algorithm - Iteration

- Apply
  - Dominant Color



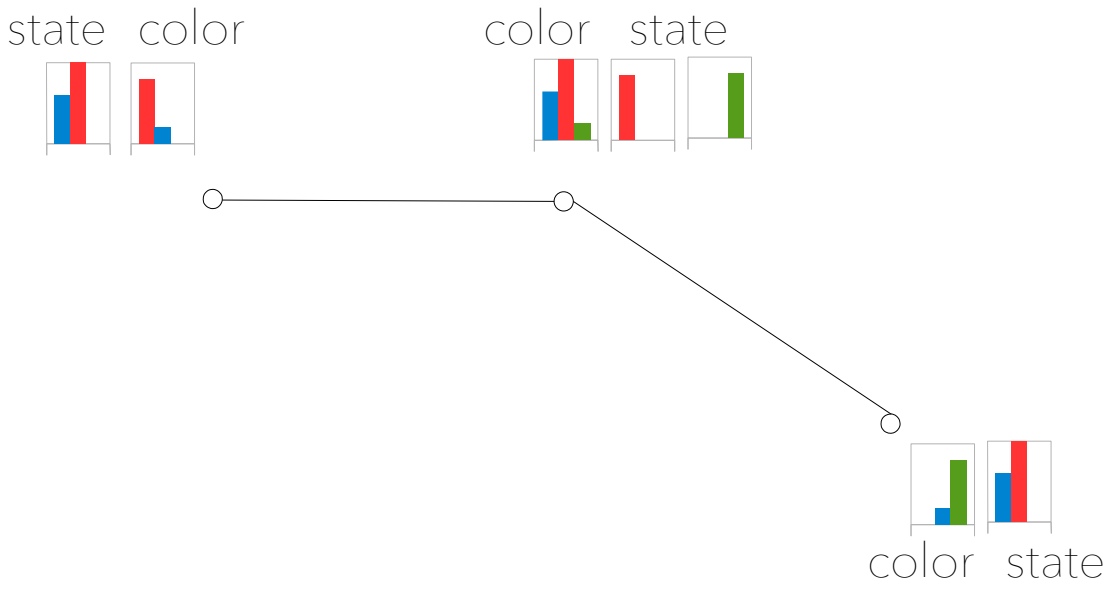
Algorithm - Iteration

- Apply
  - Dominant Color
  - Interior / Boundary

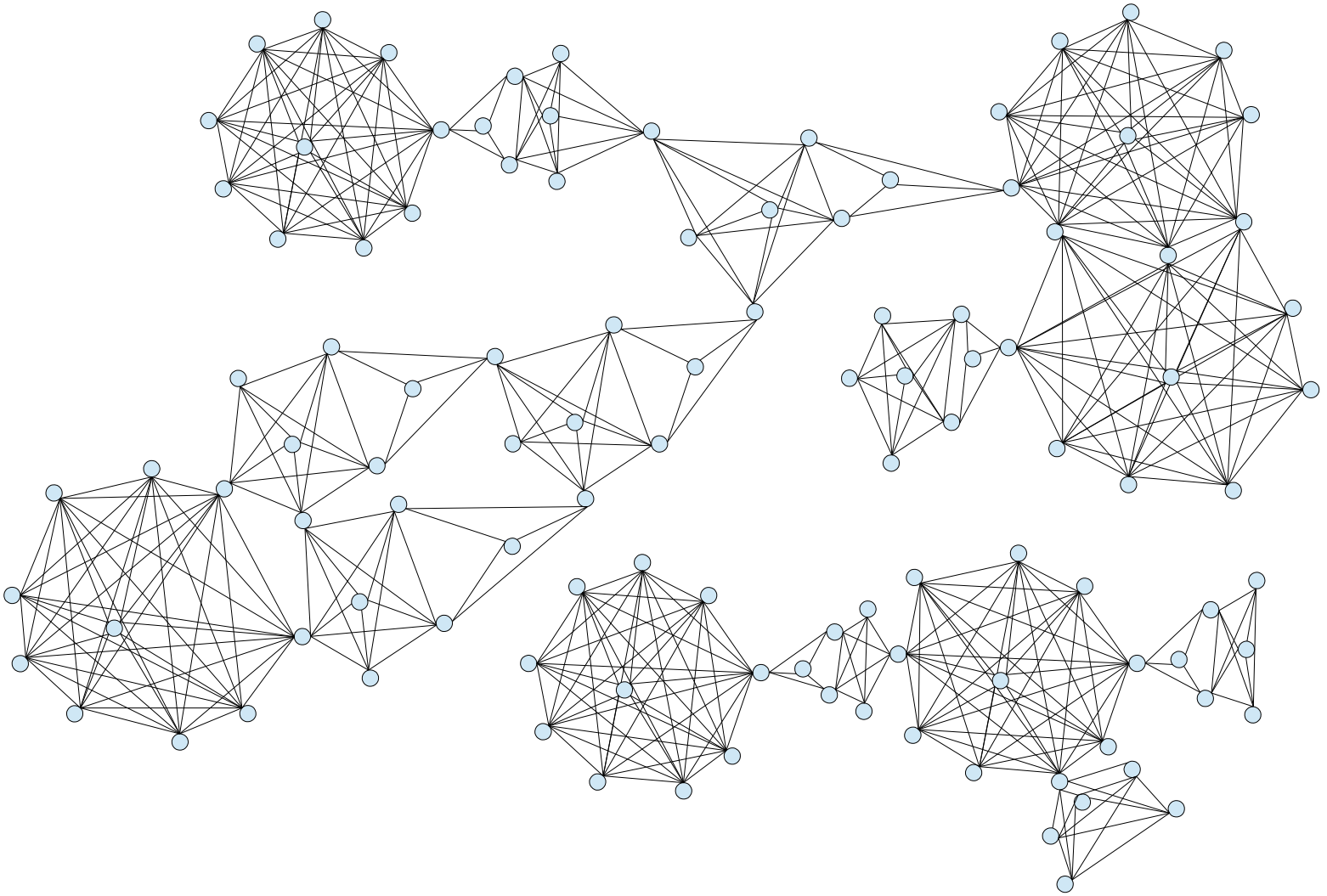


Algorithm - Iteration

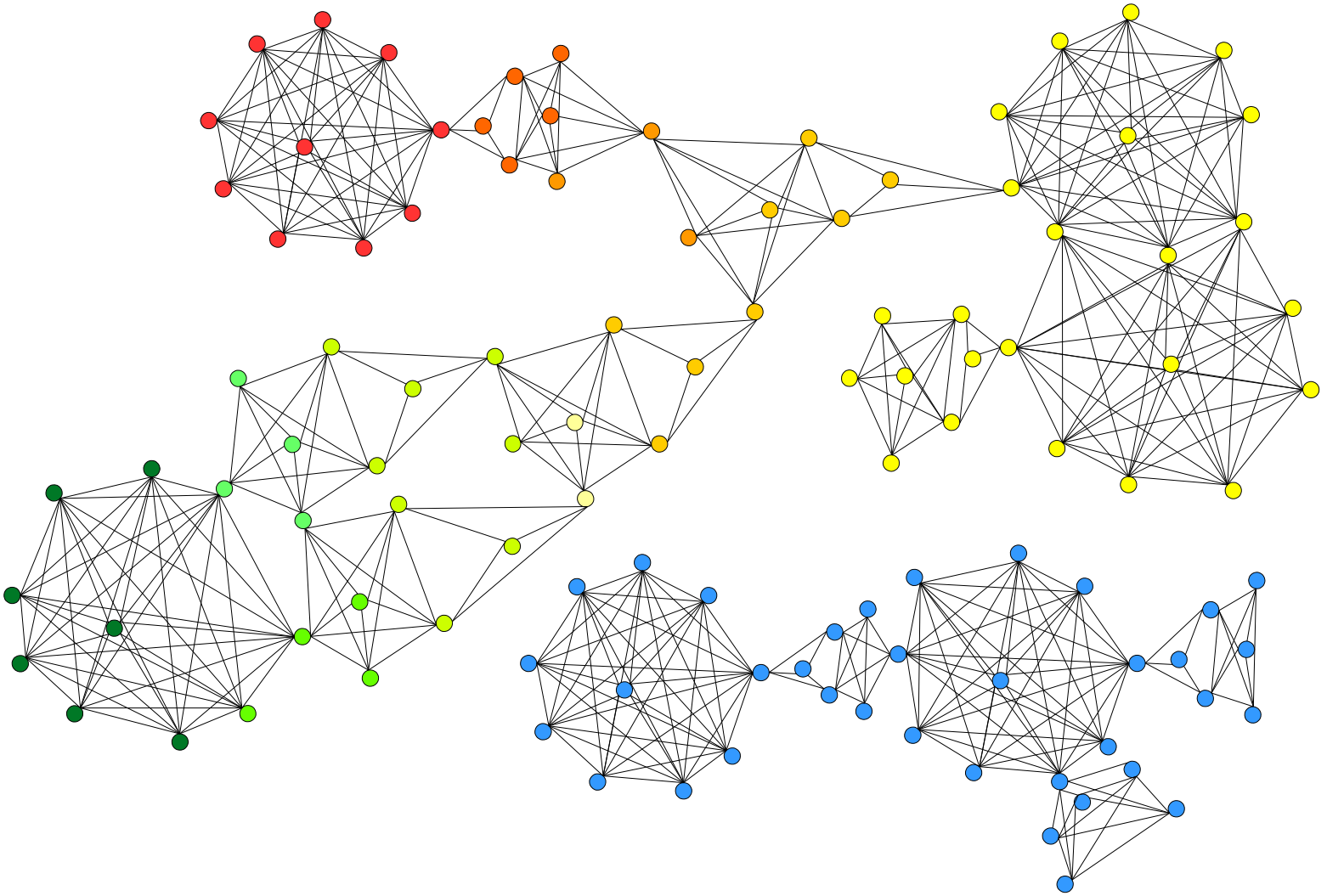
- Scatter
  - Calculation and Dissemination of neighbor's share



Algorithm

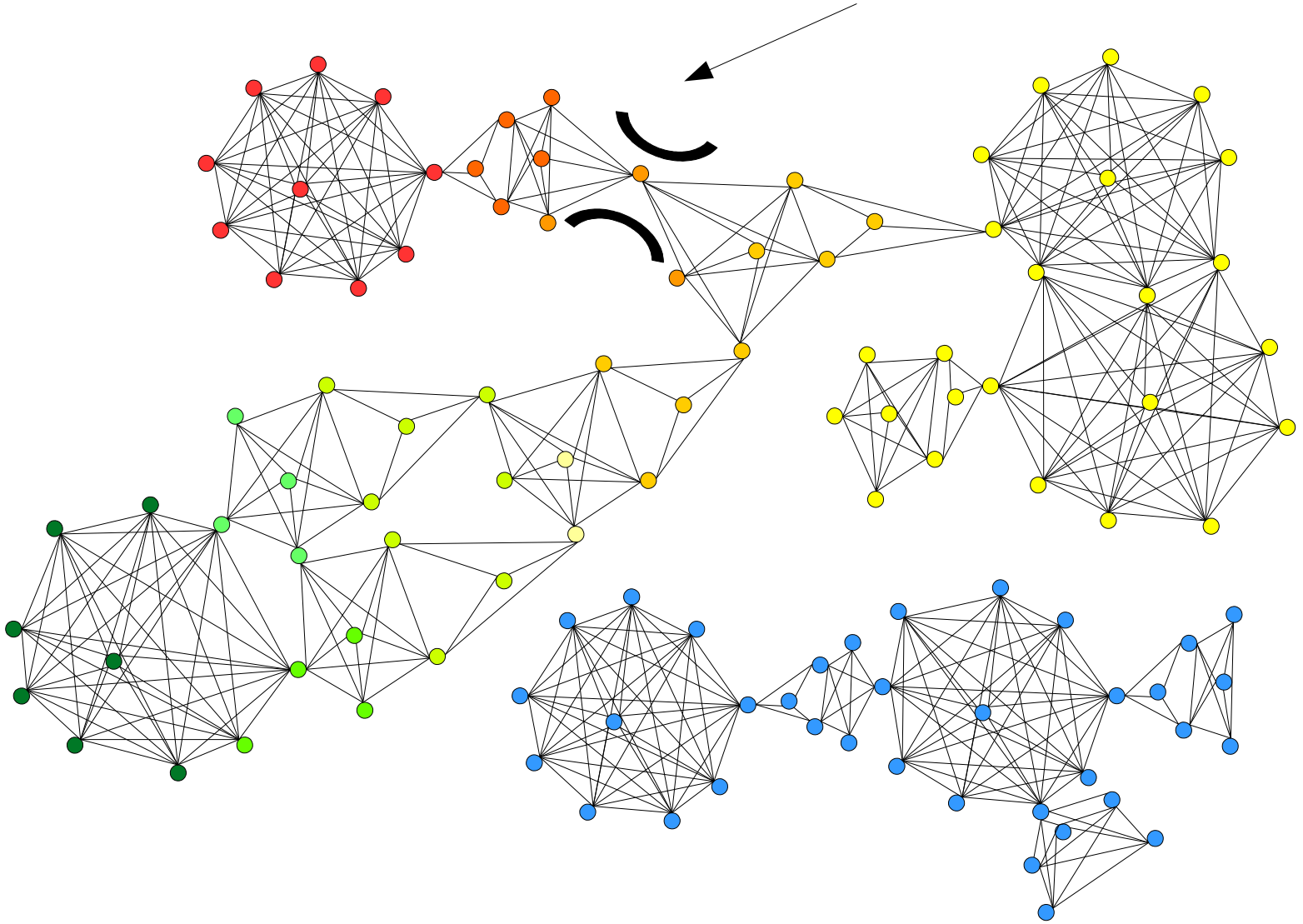


Algorithm



Algorithm

Betweenness - Number of shortest paths going through a node.



Algorithm

- Adolf Fick's first Diffusion law (1855)

  The diffusive flux is proportional to the existing concentration gradient.

$$j = -D \frac{\partial c}{\partial x}$$

$$D = Diffusivity$$

$$\frac{\partial c}{\partial x} = Concentration\ Gradient$$

1- J. Philibert Diffusion Fundamentals 2 (2005)

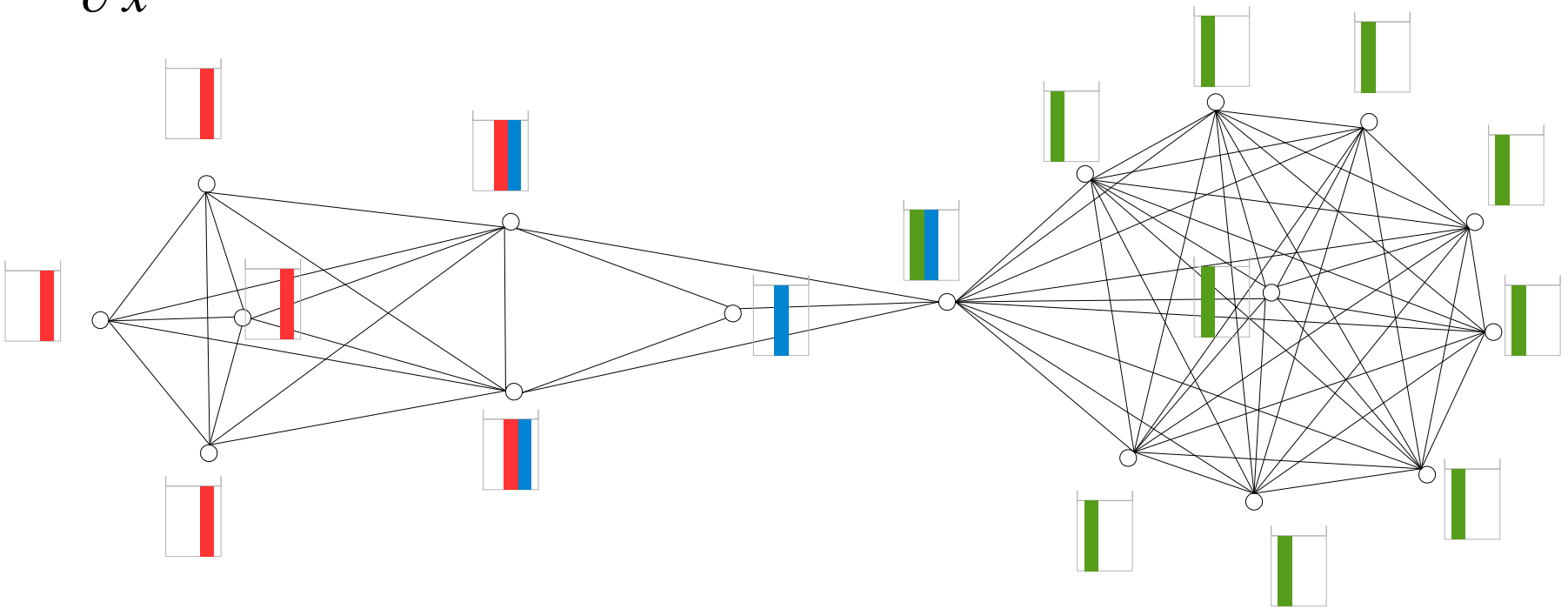
Algorithm

- Adolf Fick's Diffusion

$$j = -D\frac{\partial c}{\partial x}$$

$D = Diffusivity$

$\frac{\partial c}{\partial x} = Concentration\ Gradient$



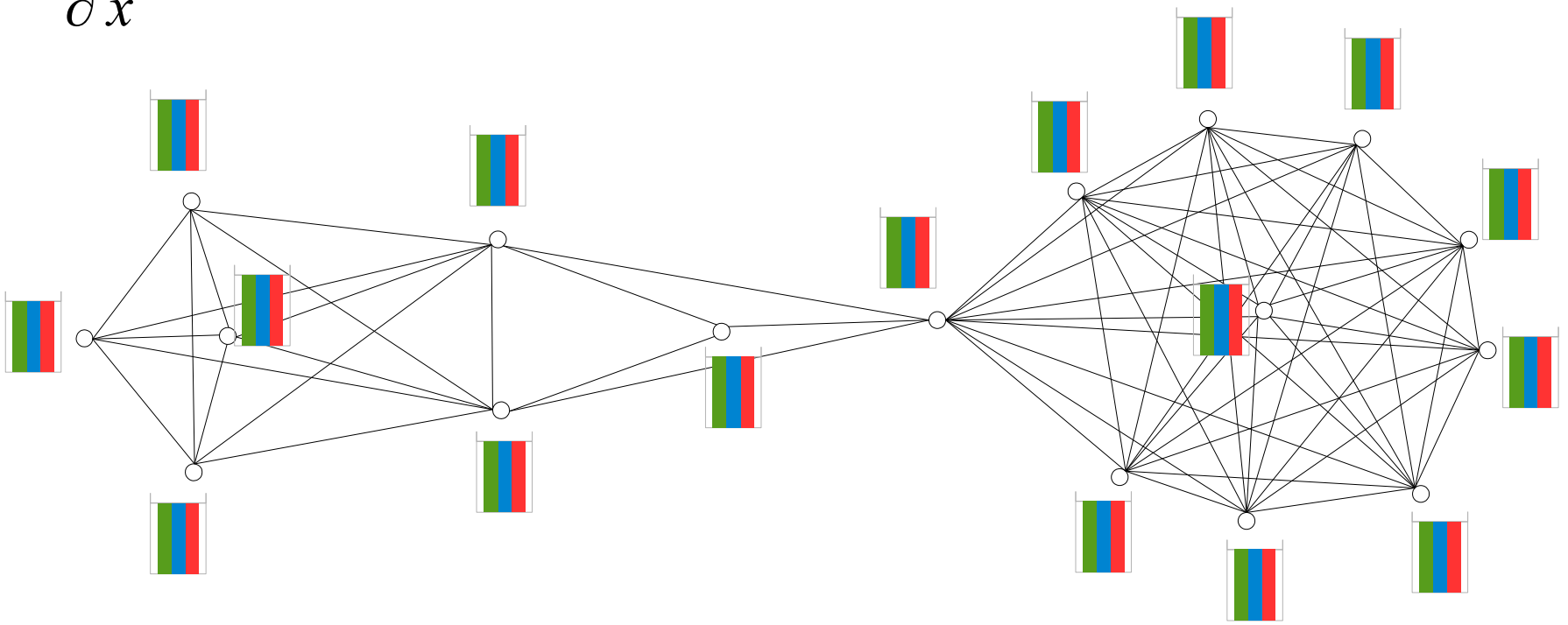
Algorithm

- Adolf Fick's Diffusion

$$j = -D\frac{\partial c}{\partial x}$$

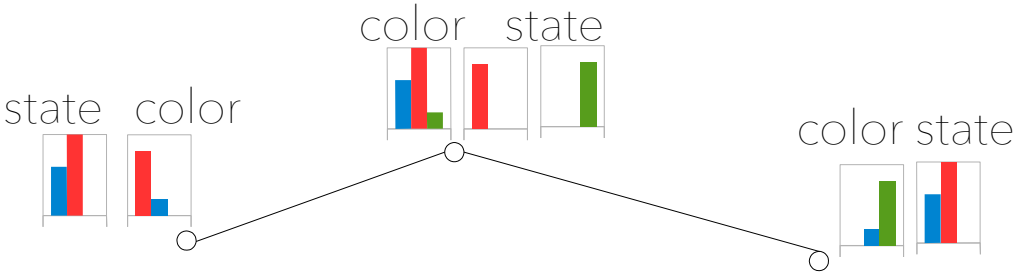
$D = Diffusivity$

$\frac{\partial c}{\partial x} = Concentration\ Gradient$



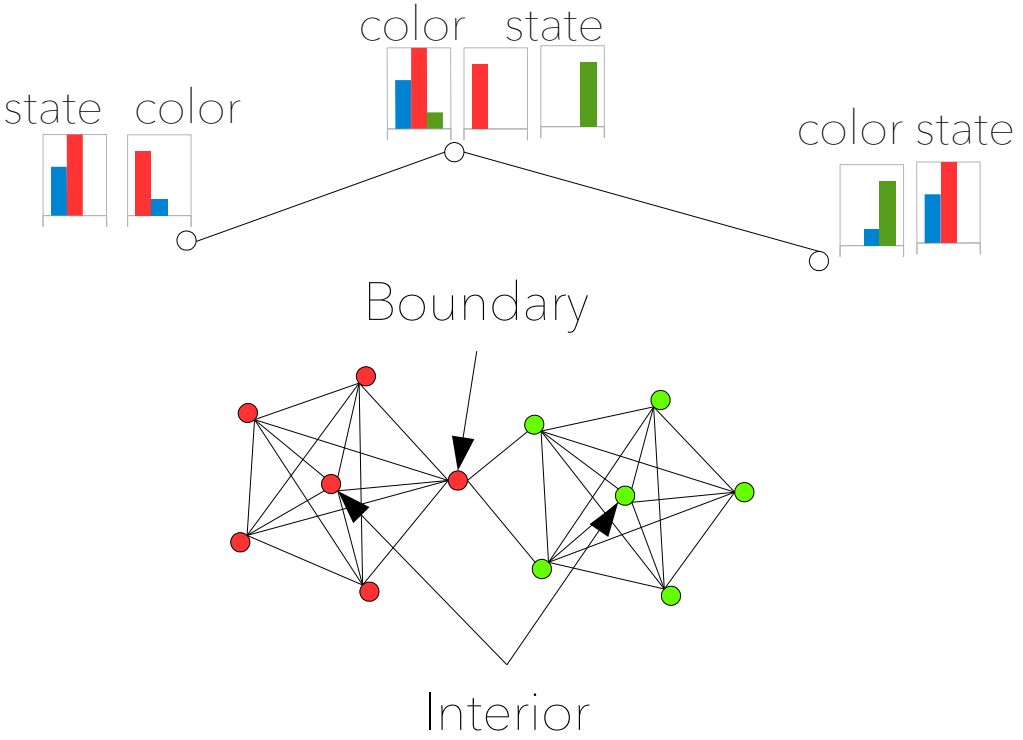
Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant



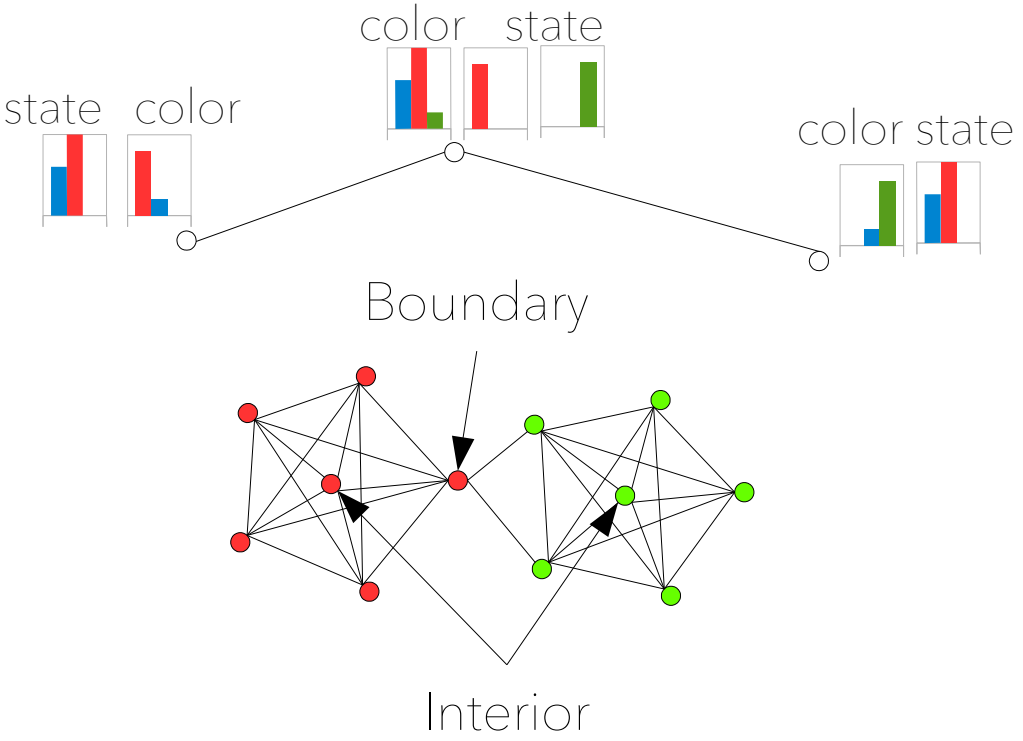
Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant
  - Interior Coefficient = α



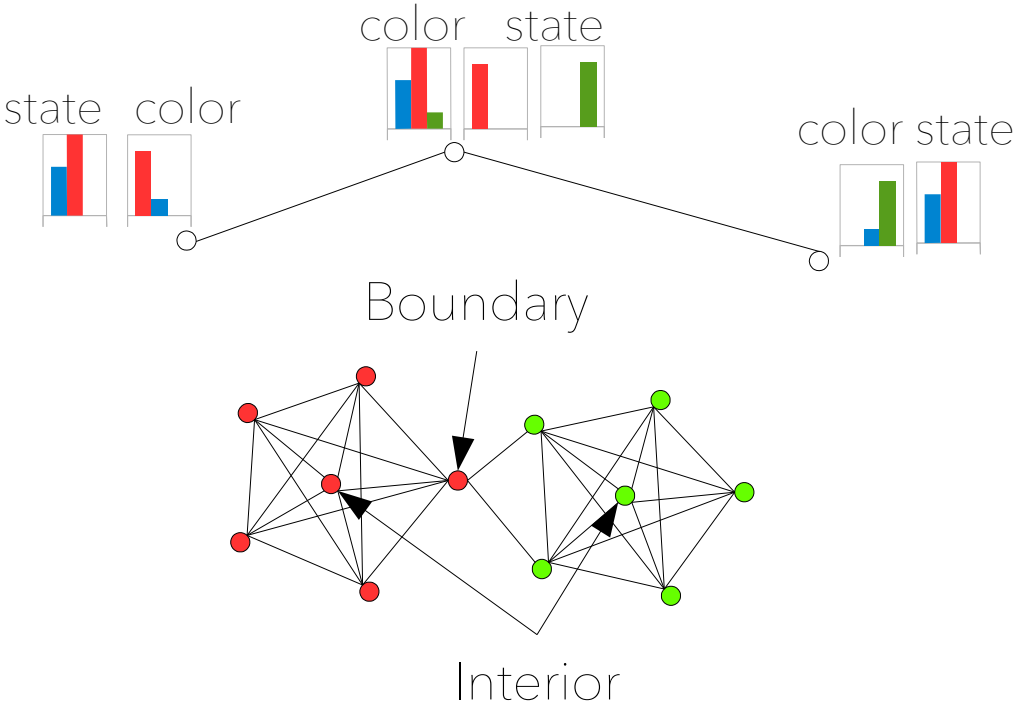
Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant
  - Interior Coefficient = α
  - Sharing portion = β



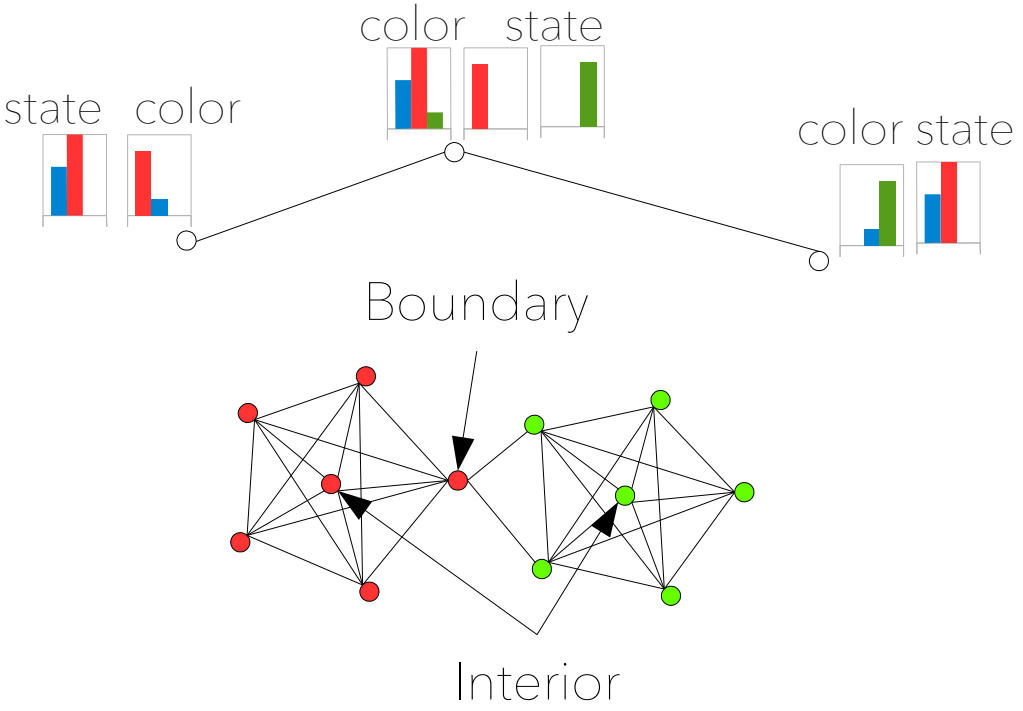
Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant
  - Interior Coefficient = α
  - Sharing portion = β
  - Bootstrapping round = γ

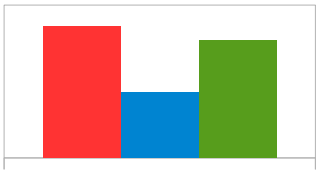


Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant
  - Interior Coefficient = α



  - Sharing portion = β
  - Bootstrapping round = γ
  - Repository



Algorithm

- Parameters
  - ColorType
    - Dominant
    - NonDominant
  - Interior Coefficient = α
  - Sharing portion = β
  - Bootstrapping round = γ

$$j = -D\frac{\partial c}{\partial x}$$

$$\frac{\partial c}{\partial x} = Concentration\ Gradient$$

$$D = Diffusivity$$

Experiments

- DataSet
  - John smith
  - 197 Documents
  - 35 True Entities
- Graph
  - 3947 Vertices
  - 362340 Edges
- BSP - Platform
  - Graphchi – C++
  - Dynamic vertex size
  - Synchronous message passing

Experiment

- Metric
  - Precision
  - Recall
  - F-Score

$$Pc = \frac{Tp}{Tp + Fp} = \frac{Number\ orrect\ Mentions\ of\ type\ C \in output}{Number\ of\ Mentions \in output}$$

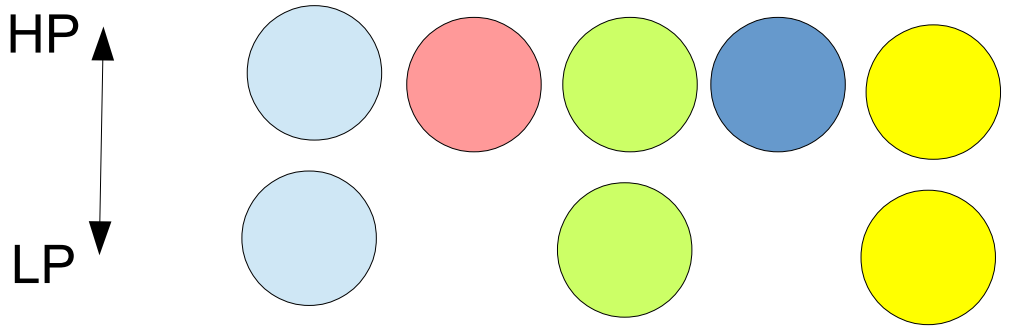
$$Rc = \frac{Tp}{Tp + Fn} = \frac{Number\ on\ correct\ Mentions\ of\ type\ C \in output}{Numbe\ of\ Mentions\ of\ type\ C}$$

$$Fc = 2 . \frac{Precision\ .\ Recall}{Precision + Recall}$$

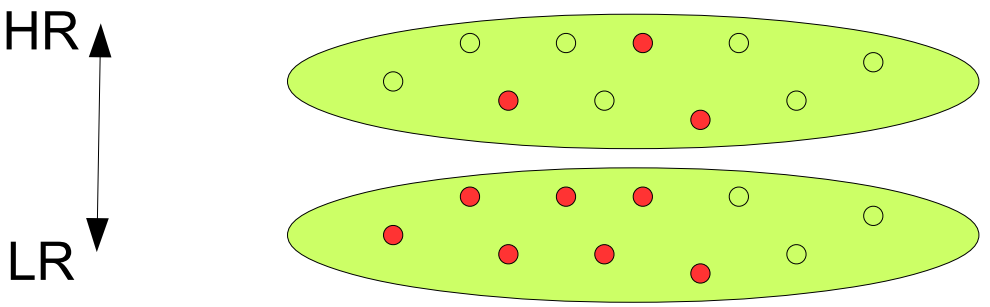
		Condition	
		Positive	Negative
Outcome	Positive	True Positive - Tp	False Positive - Tp
	Negative	False Negative - Fn	True Negative - Tn

Experiment

- Metric
  - Precision



  - Recall

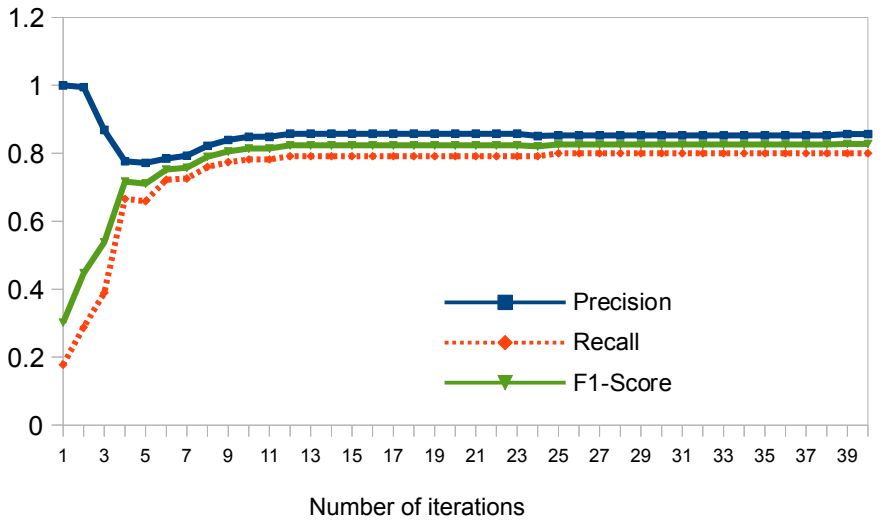


Experiment

- Parameters

	Interior	Boundary	
Dominant	Keep 96% Share 4%	Keep 4% Share 96%	
Non Dominant	Repository 100%	iteration<3	Repository 100 %
		iteration>3	Share 100 %

- Results
  - 40 rounds in 62 seconds
  - 80% F1-Score in 9 rounds
  - 82% F1-Score in 12 rounds



Experiment

	B^3 F1-Score
Bagga & Baldwin [I]	84.6 %
Rao et al. [II]	61.8 %
Google [III]	66.4 %
Our Model	83.7 %

I. Bagga & Baldwin,. 1988, Entity based Cross-Document Coreference Resolution using the Vector Space Model.

II. Rao et al.2010, Streaming Cross Document Entity Coreference Resolution.

III. Singh et al. 2010, Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:
  - Inductive

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:
  - Inductive
  - Distributed

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:
  - Inductive
  - Distributed
  - Parallel

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:
  - Inductive
  - Distributed
  - Parallel
  - Node Centric

Conclusion

- We have developed a solution for the unsupervised classification problem of community detection that is:
  - Inductive
  - Distributed
  - Parallel
  - Node Centric
  - Scaleable

Questions?