Comparing and fusing terrain network information

Bruno Gaume[⋆], Benoît Gaillard[⋆], Emmanuel Navvarro[†]

[⋆] : CLLE, Université de Toulouse II
[†] : IRIT, Université de Toulouse III

UE TAL

October 22, 2012

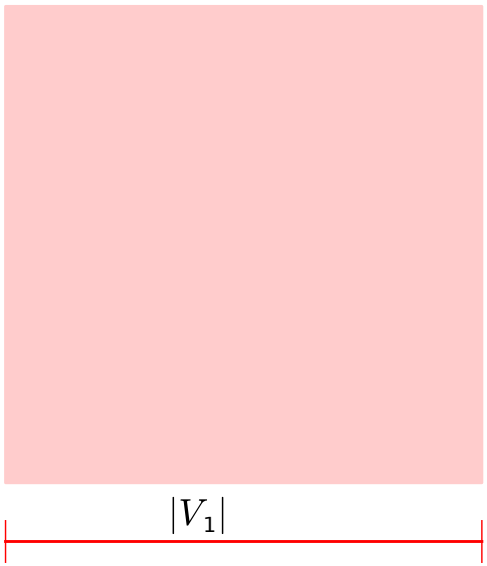
Comparing graphs' vertices

$$G_1 = (V_1, E_1), \quad G_2 = (V_2, E_2)$$

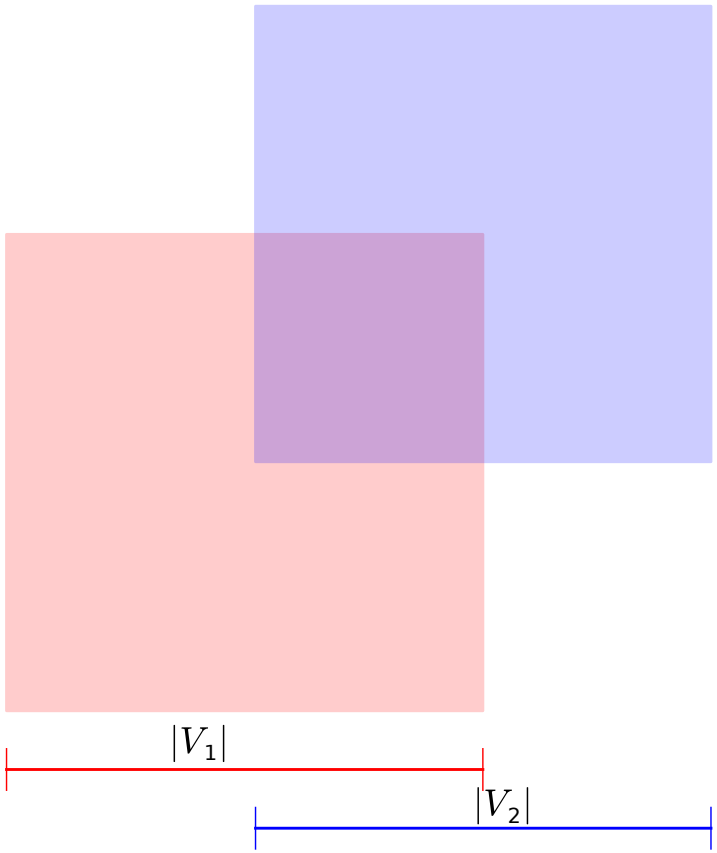
Lexical coverage :

$$R_\bullet = \frac{|V_1 \cap V_2|}{|V_2|} \qquad P_\bullet = \frac{|V_1 \cap V_2|}{|V_1|}$$

$$F_\bullet = 2.\frac{R_\bullet.P_\bullet}{R_\bullet + P_\bullet}$$



Comparing graphs' vertices



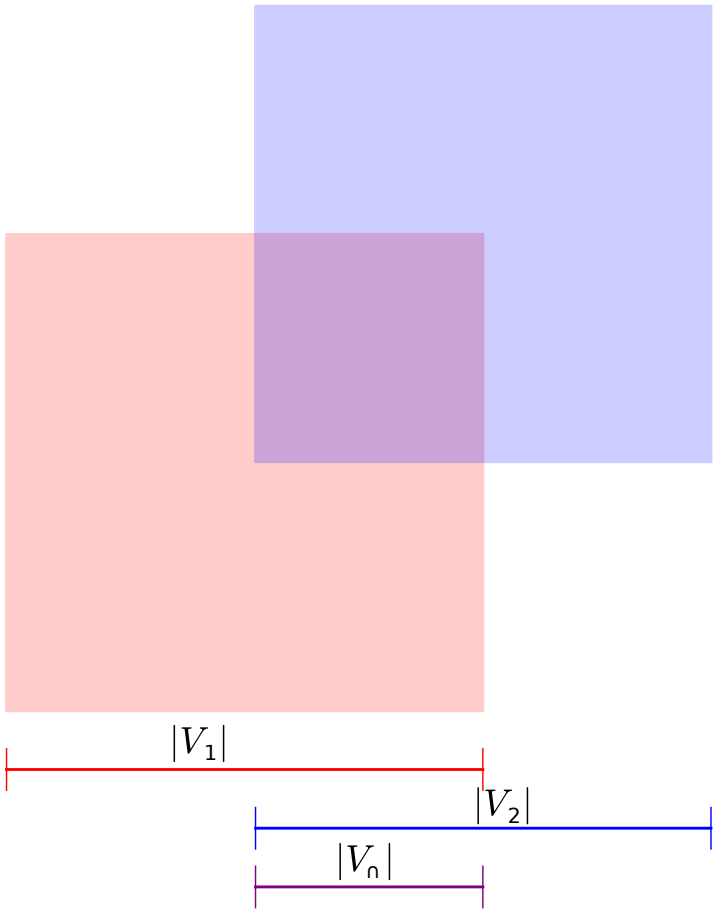
$$G_1 = (V_1, E_1), \quad G_2 = (V_2, E_2)$$

Lexical coverage :

$$R_\bullet = \frac{|V_1 \cap V_2|}{|V_2|} \qquad P_\bullet = \frac{|V_1 \cap V_2|}{|V_1|}$$

$$F_\bullet = 2.\frac{R_\bullet.P_\bullet}{R_\bullet + P_\bullet}$$

Comparing graphs' vertices



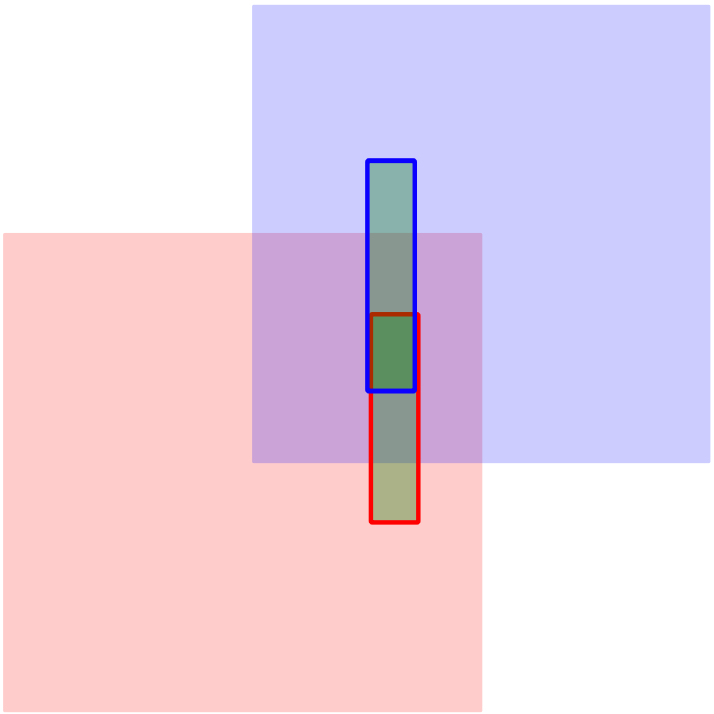
$$G_1 = (V_1, E_1), \quad G_2 = (V_2, E_2)$$

Lexical coverage :

$$R_\bullet = \frac{|V_1 \cap V_2|}{|V_2|} \qquad P_\bullet = \frac{|V_1 \cap V_2|}{|V_1|}$$

$$F_\bullet = 2.\frac{R_\bullet.P_\bullet}{R_\bullet + P_\bullet}$$

Comparing graphs' edges



Reduce graphs to common vertices,

Consider graph as a synonymy judgment,

Use **Kappa** to measure inter judge agreement:

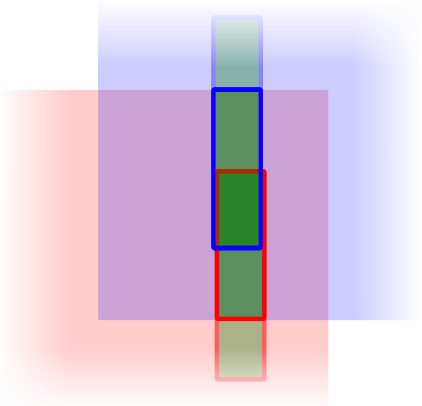
$$K_{\updownarrow}(G'_1, G'_2) = \frac{(p_0 - p_e)}{(1 - p_e)}$$

$$p_0 = \frac{1}{\omega}.(|E'_1 \cap E'_2| + |\overline{E'_1} \cap \overline{E'_2}|)$$

$$p_e = \frac{1}{\omega^2}.(|E'_1|.|E'_2| + |\overline{E'_1}|.|\overline{E'_2}|)$$

- p_0 is the proportion of observed agreement;
- p_e is the proportion of random agreement expected under the assumption of independence of decision;

Comparing graphs' edges



Reduce graphs to common vertices,

Consider graph as a synonymy judgment,

Use **Kappa** to measure inter judge agreement:

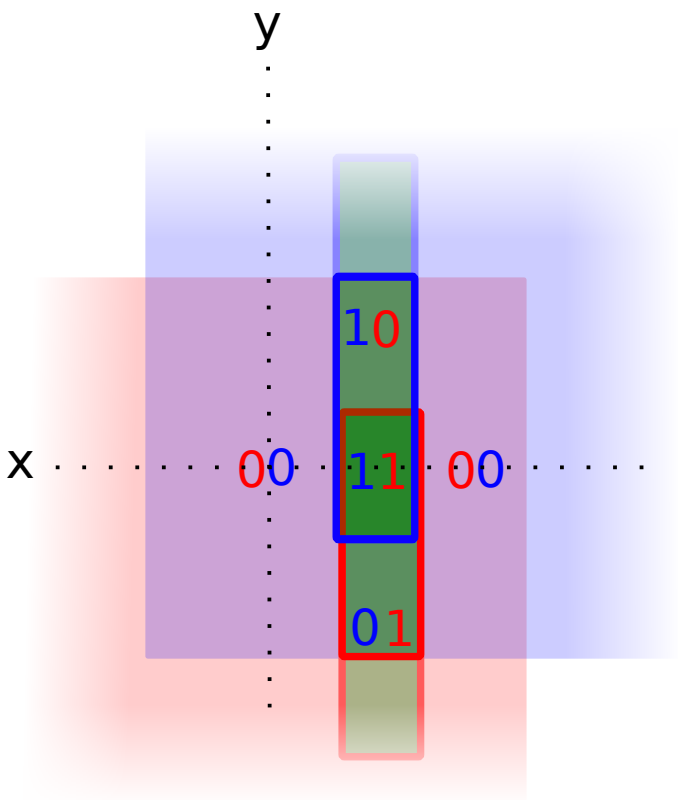
$$K_{\updownarrow}(G'_1, G'_2) = \frac{(p_0 - p_e)}{(1 - p_e)}$$

$$p_0 = \frac{1}{\omega}.(|E'_1 \cap E'_2| + |\overline{E'_1} \cap \overline{E'_2}|)$$

$$p_e = \frac{1}{\omega^2}.(|E'_1|.|E'_2| + |\overline{E'_1}|.|\overline{E'_2}|)$$

- p_0 is the proportion of observed agreement;
- p_e is the proportion of random agreement expected under the assumption of independence of decision;

Comparing graphs' edges



Reduce graphs to common vertices,

Consider graph as a synonymy judgment,

Use **Kappa** to measure inter judge agreement:

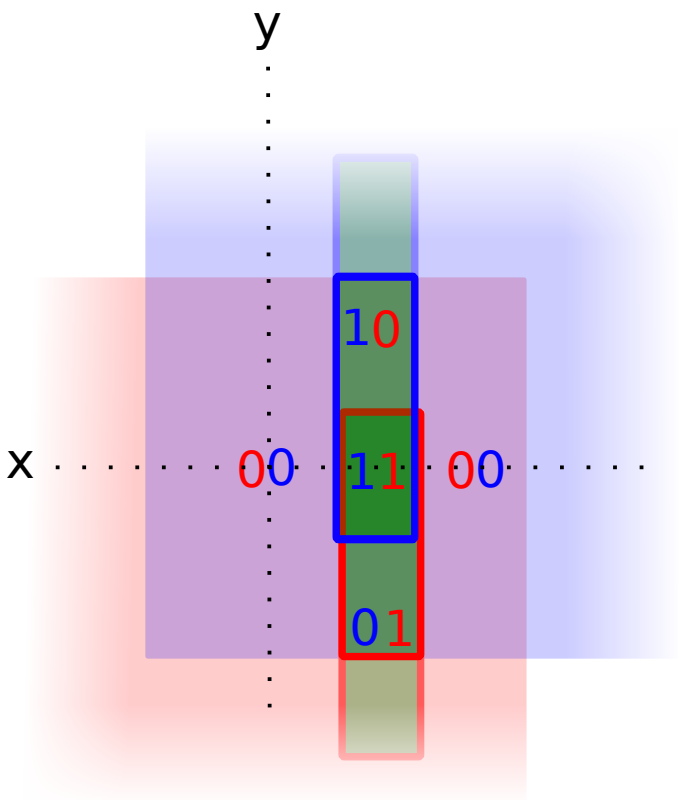
$$K_{\updownarrow}(G'_1, G'_2) = \frac{(p_0 - p_e)}{(1 - p_e)}$$

$$p_0 = \frac{1}{\omega}.(|E'_1 \cap E'_2| + |\overline{E'_1} \cap \overline{E'_2}|)$$

$$p_e = \frac{1}{\omega^2}.(|E'_1|.|E'_2| + |\overline{E'_1}|.|\overline{E'_2}|)$$

- p_0 is the proportion of observed agreement;
- p_e is the proportion of random agreement expected under the assumption of independence of decision;

Comparing graphs' edges



Reduce graphs to common vertices,

Consider graph as a synonymy judgment,

Use **Kappa** to measure inter judge agreement:

$$K_{\updownarrow}(G'_1, G'_2) = \frac{(p_0 - p_e)}{(1 - p_e)}$$

$$p_0 = \frac{1}{\omega}.(|E'_1 \cap E'_2| + |\overline{E'_1} \cap \overline{E'_2}|)$$

$$p_e = \frac{1}{\omega^2}.(|E'_1|.|E'_2| + |\overline{E'_1}|.|\overline{E'_2}|)$$

- p_0 is the proportion of observed agreement;
- p_e is the proportion of random agreement expected under the assumption of independence of decision;

Experiment: Comparison of 7 synonymy resources

5 French dictionaries:

(General purpose, paper dictionaries)

- Bailly
- Benac
- Bertaud
- Larousse
- Robert

2 English resources:

- Wiktionary
- Princeton Wordnet

	n	m	$\langle k \rangle$	n_{lcc}	m_{lcc}	C	L_{lcc}	λ	r^2
Bai_V	3082	3648	2.46	2774	3417	0.04	8.24	-2.33	0.94
Ben_V	3549	4680	2.73	3318	4528	0.03	6.52	-2.10	0.96
Ber_V	6561	25177	7.71	6524	25149	0.13	4.52	-1.88	0.93
Lar_V	5377	22042	8.44	5193	21926	0.17	4.61	-1.94	0.88
Rob_V	7357	26567	7.48	7056	26401	0.12	4.59	-2.01	0.93
PWN_V	11529	23019	6.3	6534	20806	0.47	5.9	-2.4	0.90
Wik_V	7339	8353	2.8	4285	6093	0.11	8.9	-2.4	0.94

Results: A Weak Agreement ...

$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	0.583	0.309	0.255	0.288	
Ben_V		0.389	0.276	0.293	
Ber_V			0.416	0.538	
Lar_V				0.518	
PWN_V					0.247

Results: A Weak Agreement ...

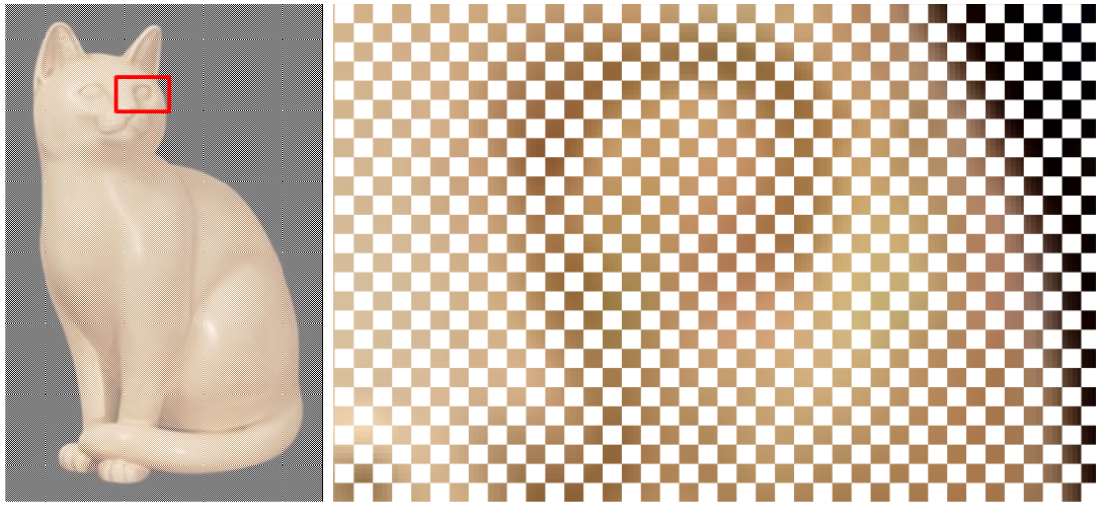
$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	0.583	0.309	0.255	0.288	
Ben_V		0.389	0.276	0.293	
Ber_V			0.416	0.538	
Lar_V				0.518	
PWN_V					0.247

What's wrong ??

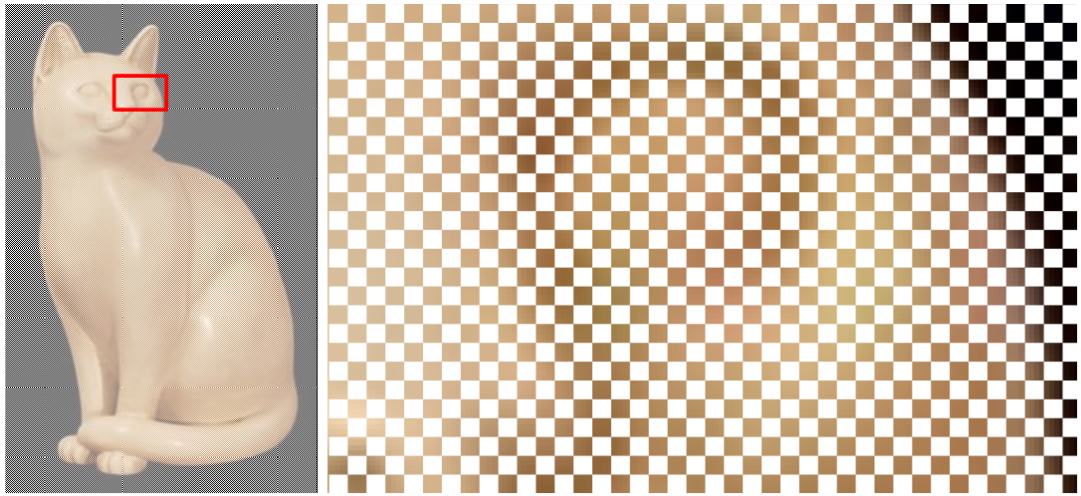
- Why do resources describing the same lexicon appear so different ?

The picture metaphor (1/2)

The picture of a cat



Each **odd** pixel painted in white...



Each **even** pixel painted in white...

The picture metaphor (2/2)

$Even =$ 

The picture metaphor (2/2)

$Even =$ 

$Odd =$ 

Can you see a difference ?

however... $sim(A, B) \approx 0$, when computed at **pixel level**.

Similarly on graphs: take a step back !

Weak agreement at the **edge level**

but

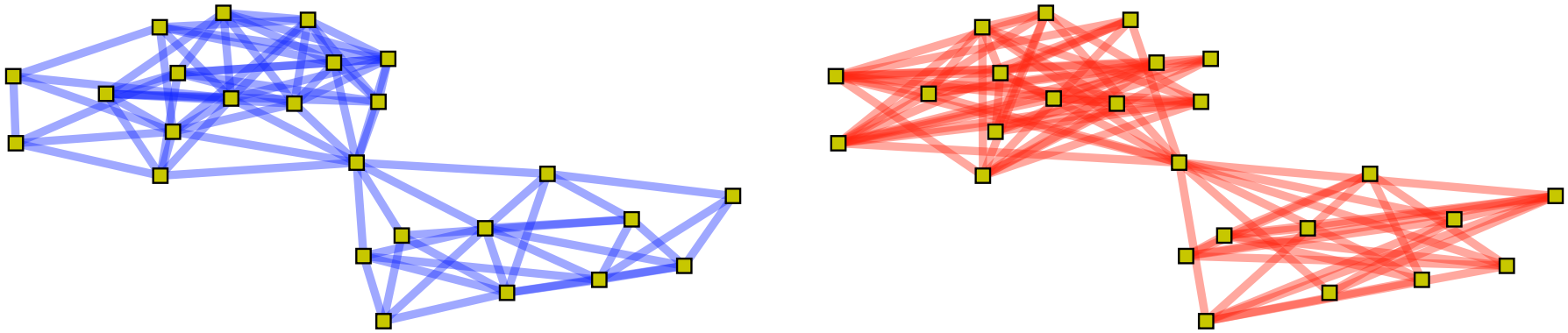
a stronger agreement at a **coarser grain level**...

Similarly on graphs: take a step back !

Weak agreement at the **edge level**

but

a stronger agreement at a **coarser grain level**...

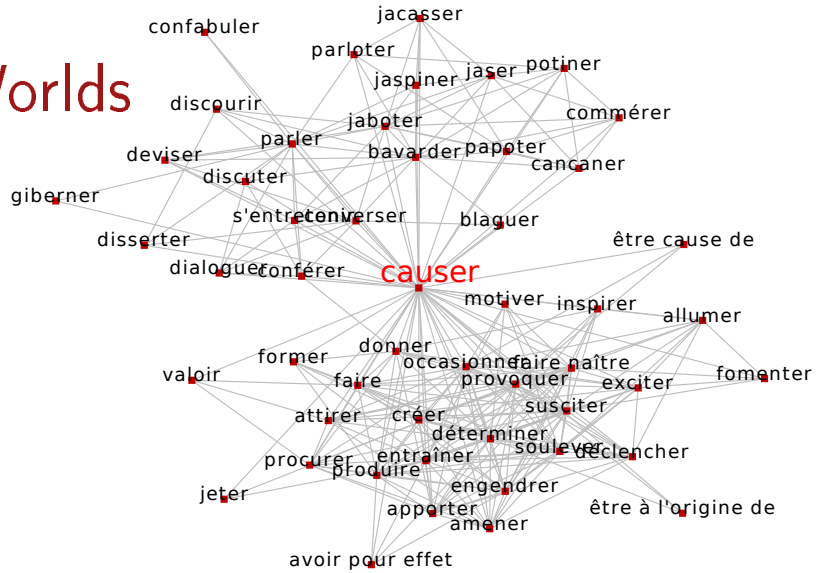


As for the pictures: **no edge in common** between these two graphs.

How can we look at our graph at different grain levels ?

Lexical Networks: Hierarchical Small Worlds

- Low density
- Short paths
- Heavy tailed degree distribution
- High clustering coefficient: dense zones



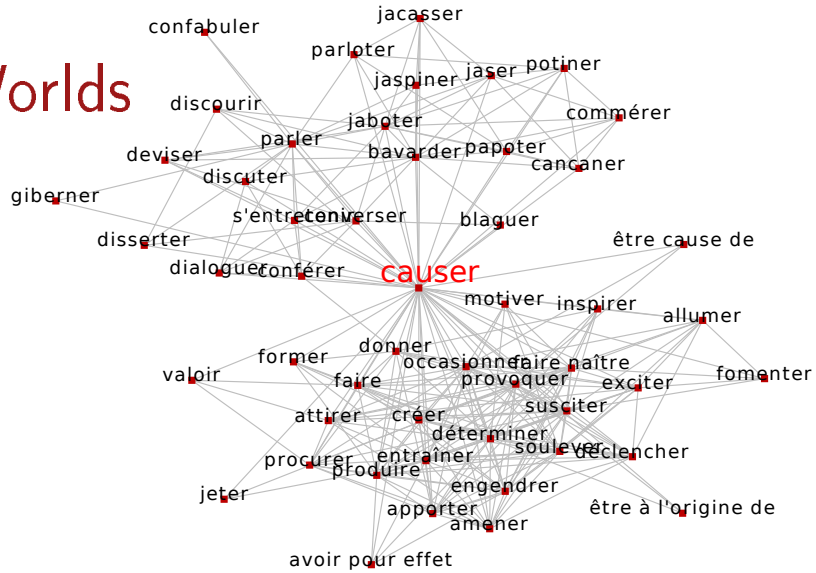
Random Walks

- **Idea**: if (u, v) are **in the same *"cluster"***, they may **not** be **adjacent**, but **many shorth paths** lead from u to v.
- Random walkers tend to be trapped into clusters,
- Note: possible approach for clustering

How can we look at our graph at different grain levels ?

Lexical Networks: Hierarchical Small Worlds

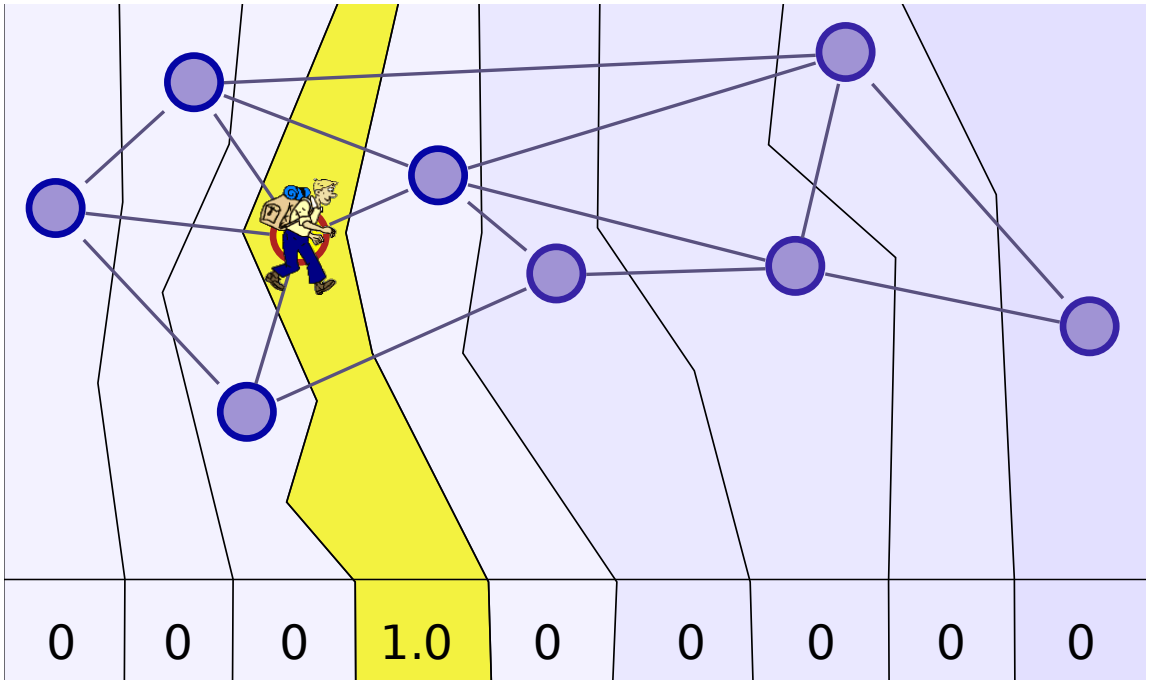
- Low density
- Short paths
- Heavy tailed degree distribution
- High clustering coefficient: dense zones



Random Walks

- **Idea**: if (u, v) are **in the same *"cluster"***, they may **not** be **adjacent**, but **many shorth paths** lead from u to v.
- Random walkers tend to be trapped into clusters,
- Note: possible approach for clustering

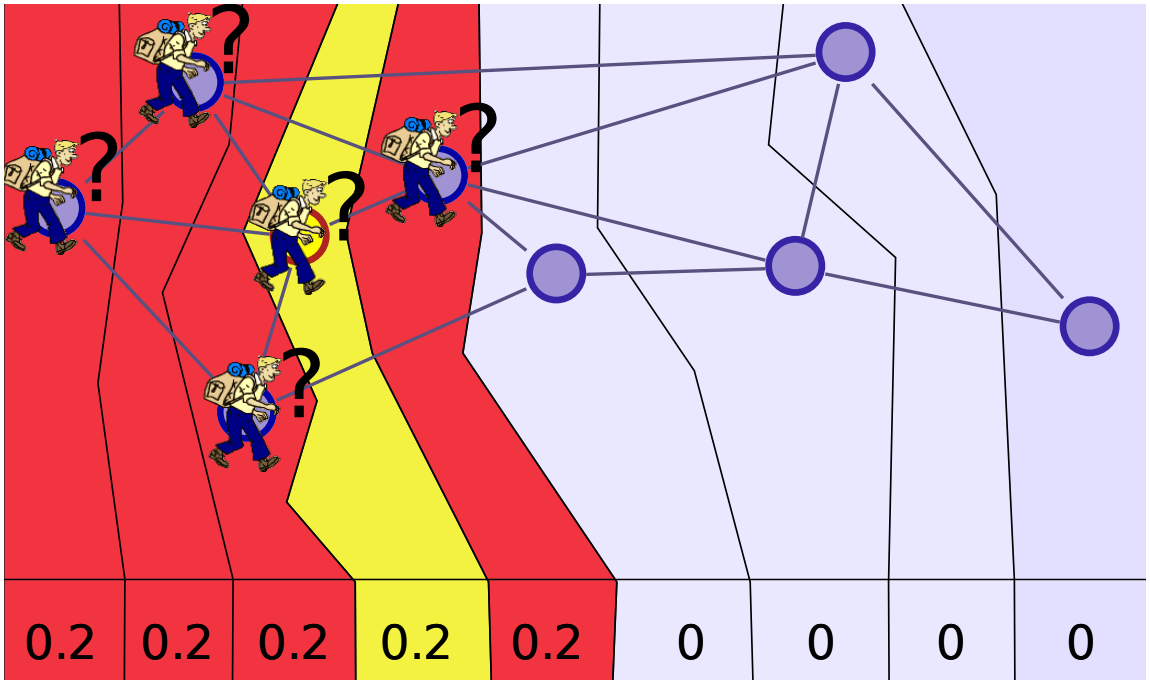
Random walks on graph...



1. start from a node u,
2. walk to a neighbour with equal probability,
3. walk to a neighbour with equal probability,
4. etc...

$$t = 0 \ , \quad P^t(u, *) = [0, 0, 0, 1.0, 0, 0, 0, 0, 0]$$

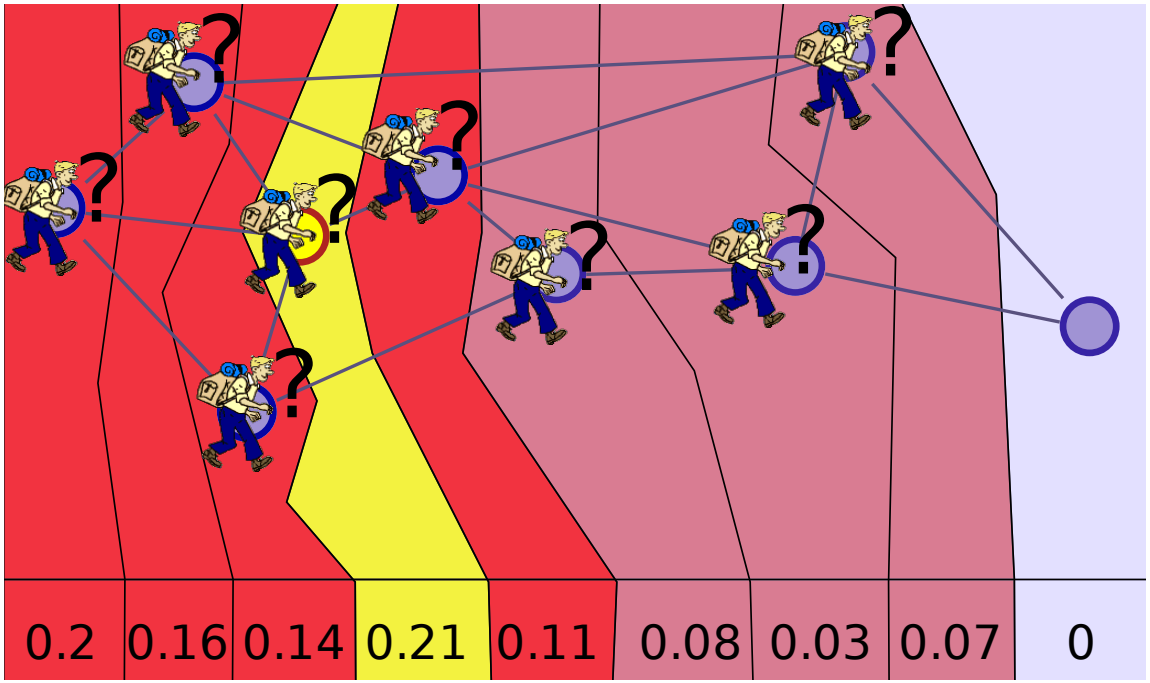
Random walks on graph...



1. start from a node u,
2. walk to a neighbour with equal probability,
3. walk to a neighbour with equal probability,
4. etc...

$$t = 1 \ , \quad P^t(u, *) = [0.2, 0.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0]$$

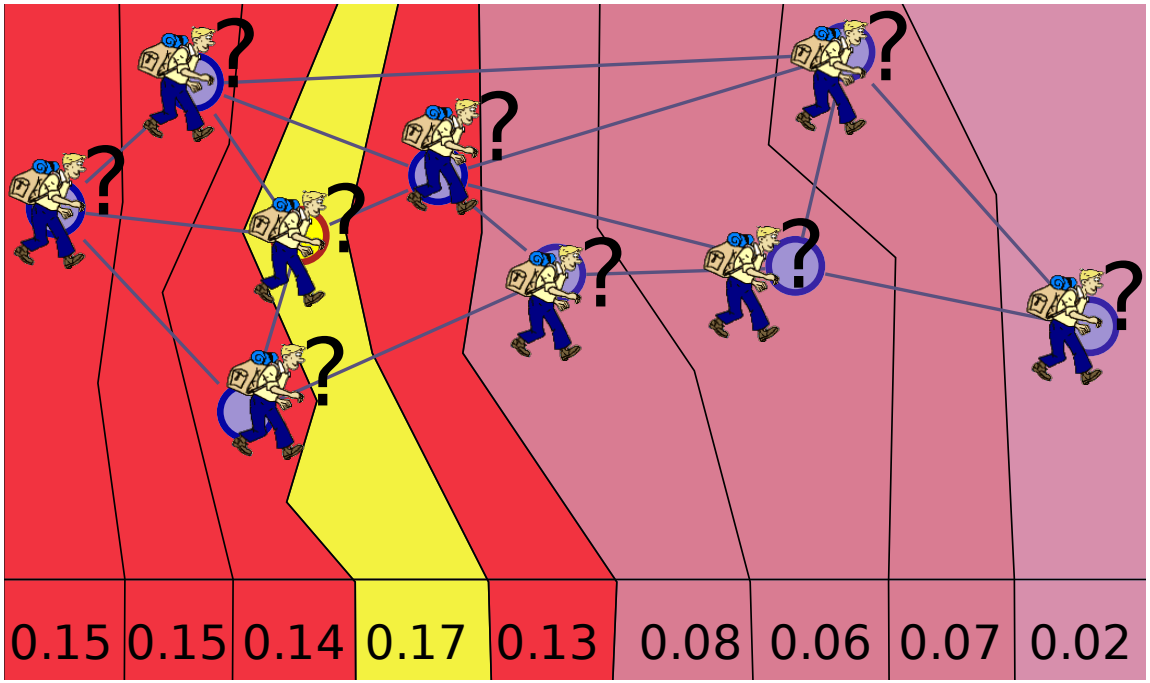
Random walks on graph...



1. start from a node u,
2. walk to a neighbour with equal probability,
3. walk to a neighbour with equal probability,
4. etc...

$$t = 2 \ , \quad P^t(u, *) = [0.2, 0.16, 0.14, 0.21, 0.11, 0.08, 0.03, 0.07, 0]$$

Random walks on graph...



1. start from a node u,
2. walk to a neighbour with equal probability,
3. walk to a neighbour with equal probability,
4. etc...

$$t = 3\ , \quad P^t(u, *) = [0.15, 0.15, 0.14, 0.17, 0.13, 0.08, 0.06, 0.07, 0.02]$$

Strong and weak confluence

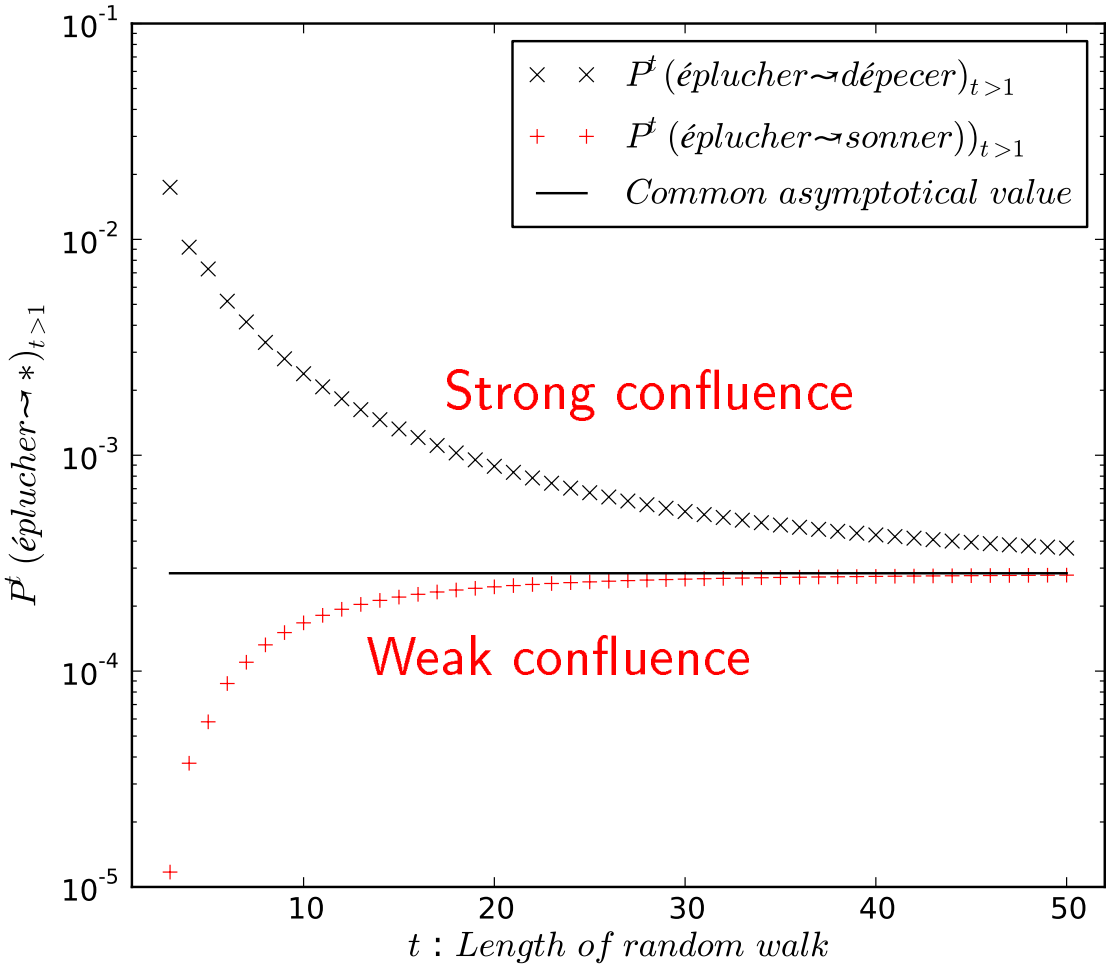
- **Long walks**: probability of reaching a node v only depends on v's degree:

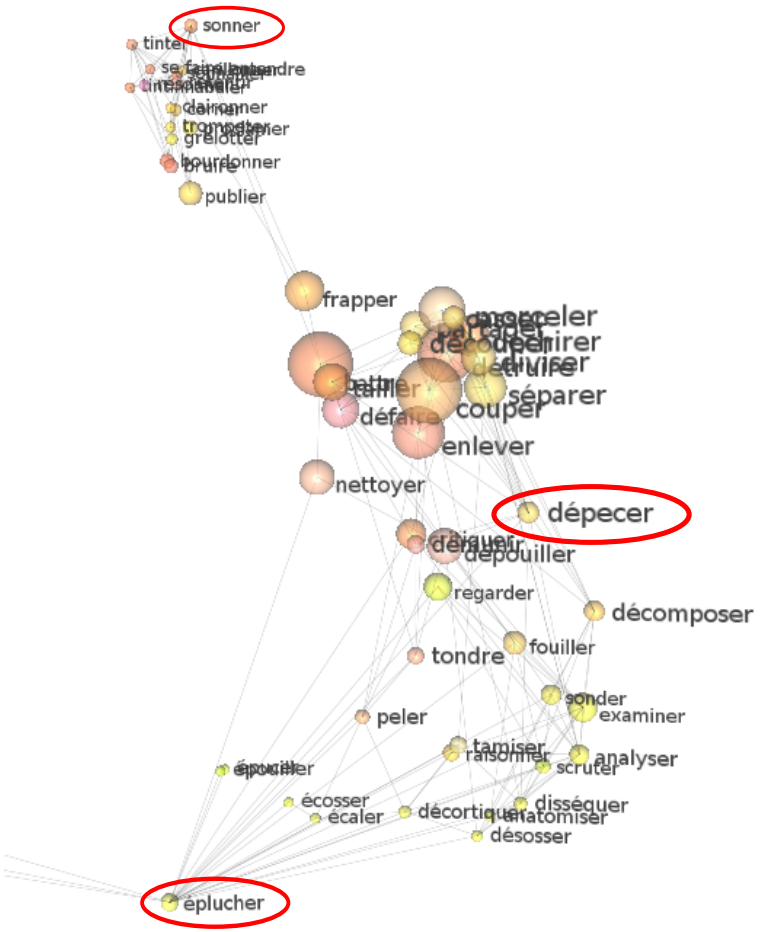
$$\lim_{t \to \infty} P^t(u, v) = \frac{deg(v)}{\sum_{n \in V} deg(n)} = \pi_v \quad (1)$$

- **Short walks**: high probability of staying in a dense zones:
  - $P_t(u, v) > \pi_v$ if u et v in the same cluster: *strong confluence*
  - $P_t(u, v) < \pi_v$ otherwise: *weak confluence*

Illustration of weak and strong confluence (1/2)

Confluences, in Rob_V, **between:**
éplucher (peel) $\leftrightarrow$ *dépecer (tear apart)* **and** *éplucher (peel)* $\nleftrightarrow$ *sonner (ring)*.





http://prox.irit.fr

Illustration of weak and strong confluence (1/2)

In Robert:

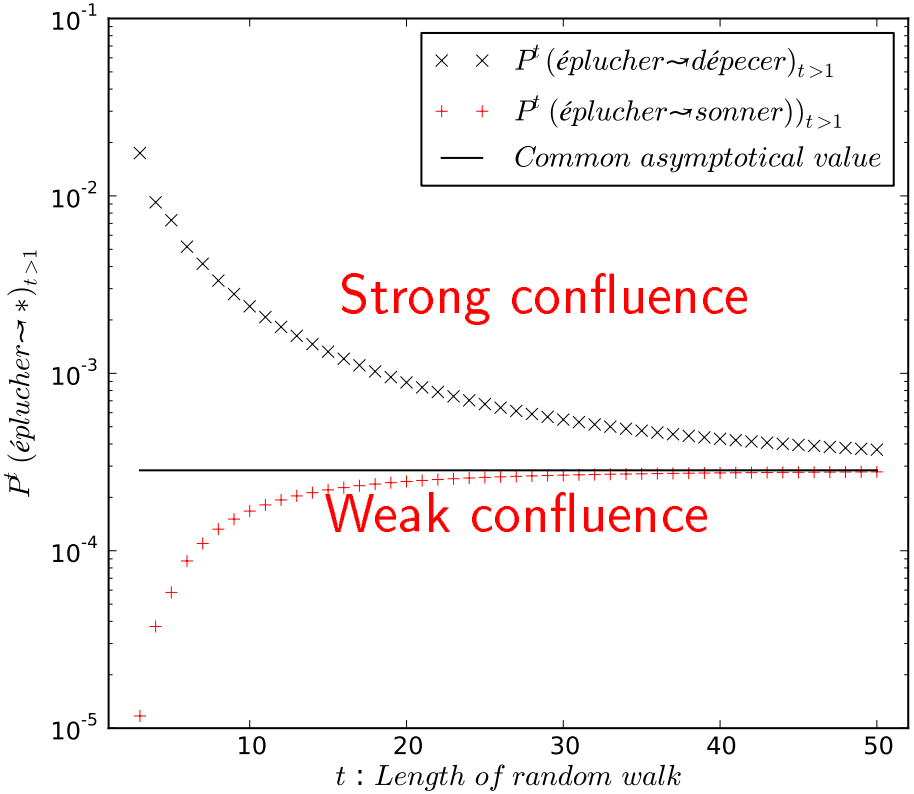
éplucher (peel) $\leftrightarrow$ *dépecer (tear apart)*

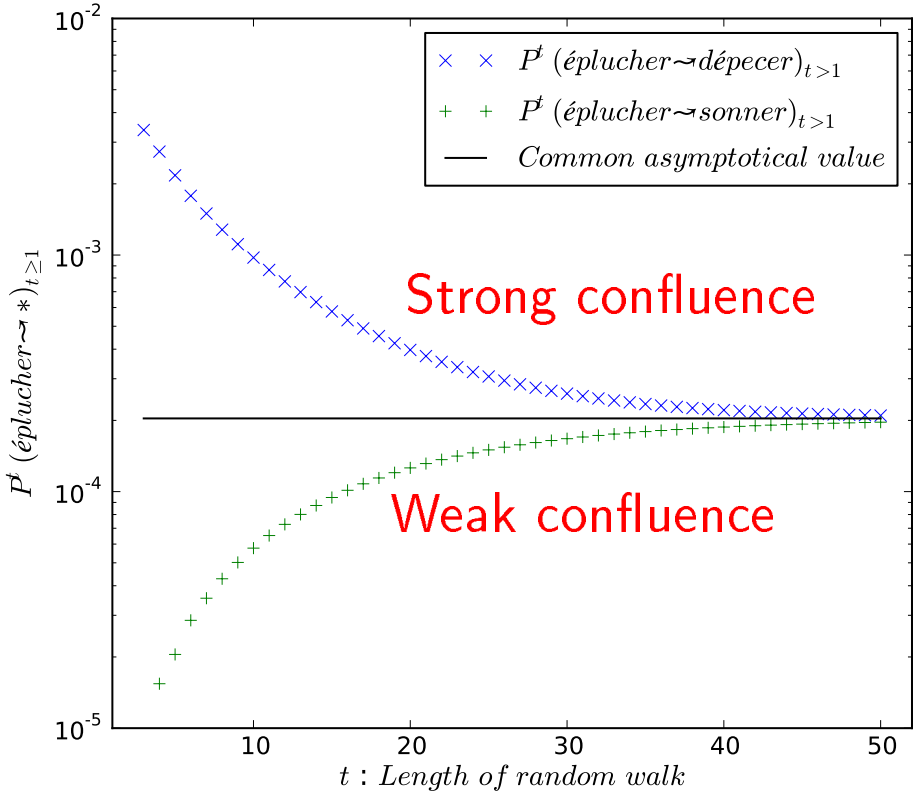
éplucher (peel) $\not\leftrightarrow$ *sonner (ring)*

In Larousse:

éplucher (peel) $\not\leftrightarrow$ *dépecer (tear apart)*

éplucher (peel) $\not\leftrightarrow$ *sonner (ring)*

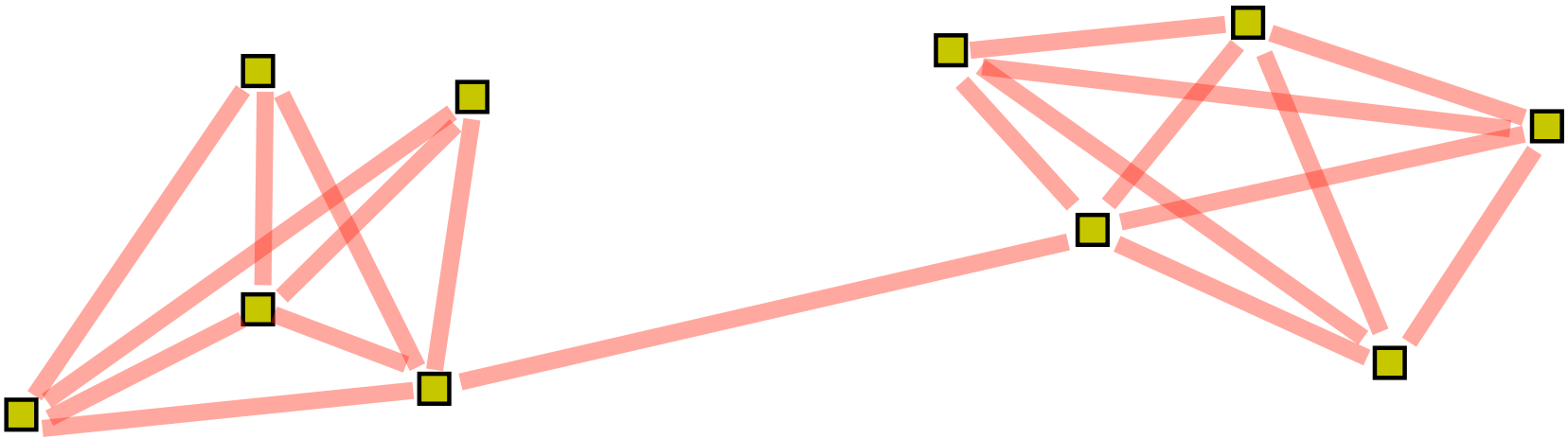




Signature of confluence

4 types of pairs of vertices:

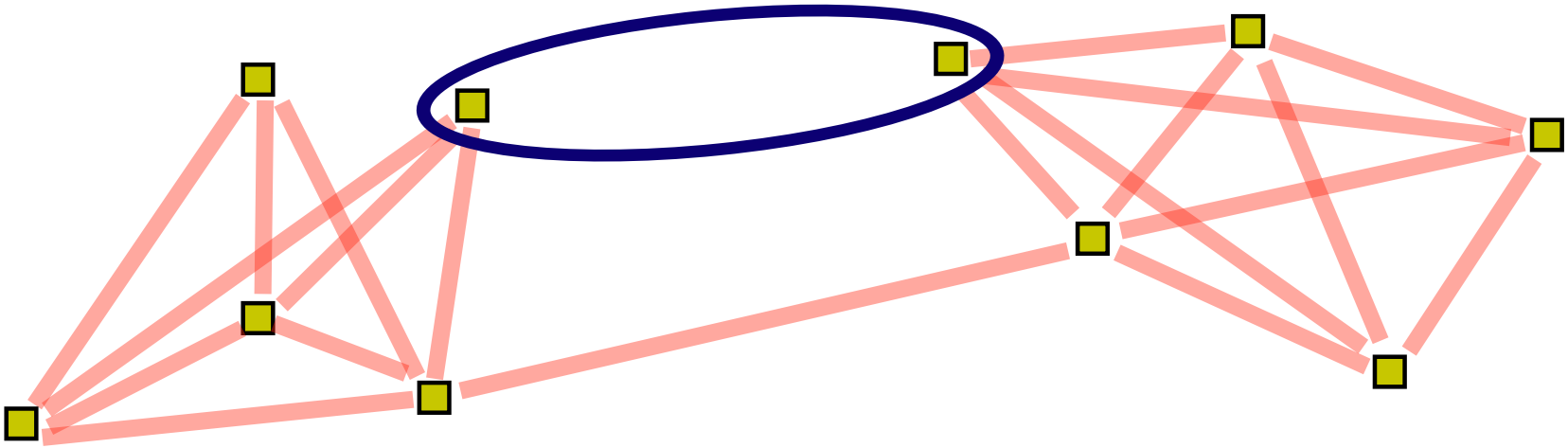
Synonymy	Confluence	
No	weak	Unrelated
No	strong	Potential synonyms
Yes	weak	Shortcut
Yes	strong	Strong synonyms



Signature of confluence

4 types of pairs of vertices:

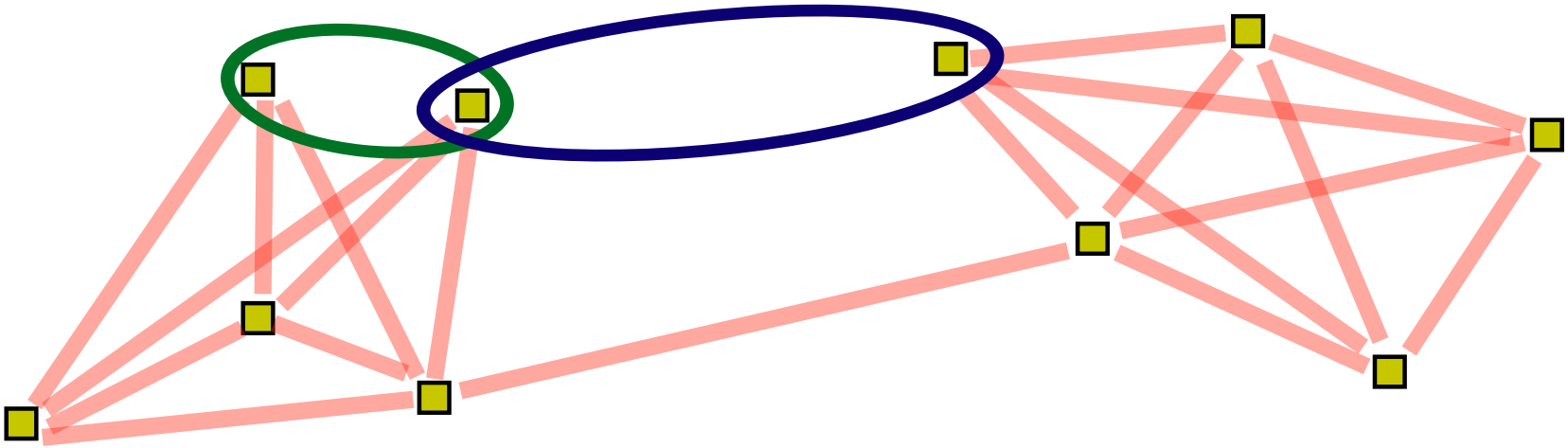
Synonymy	Confluence	
No	weak	Unrelated
No	strong	Potential synonyms
Yes	weak	Shortcut
Yes	strong	Strong synonyms



Signature of confluence

4 types of pairs of vertices:

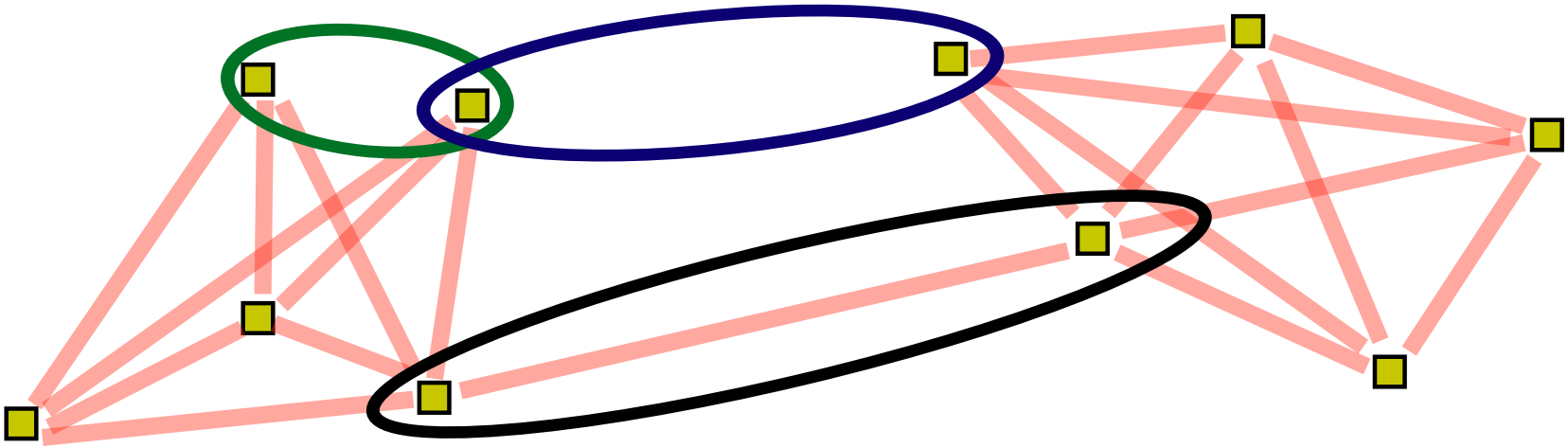
Synonymy	Confluence	
No	weak	Unrelated
No	strong	Potential synonyms
Yes	weak	Shortcut
Yes	strong	Strong synonyms



Signature of confluence

4 types of pairs of vertices:

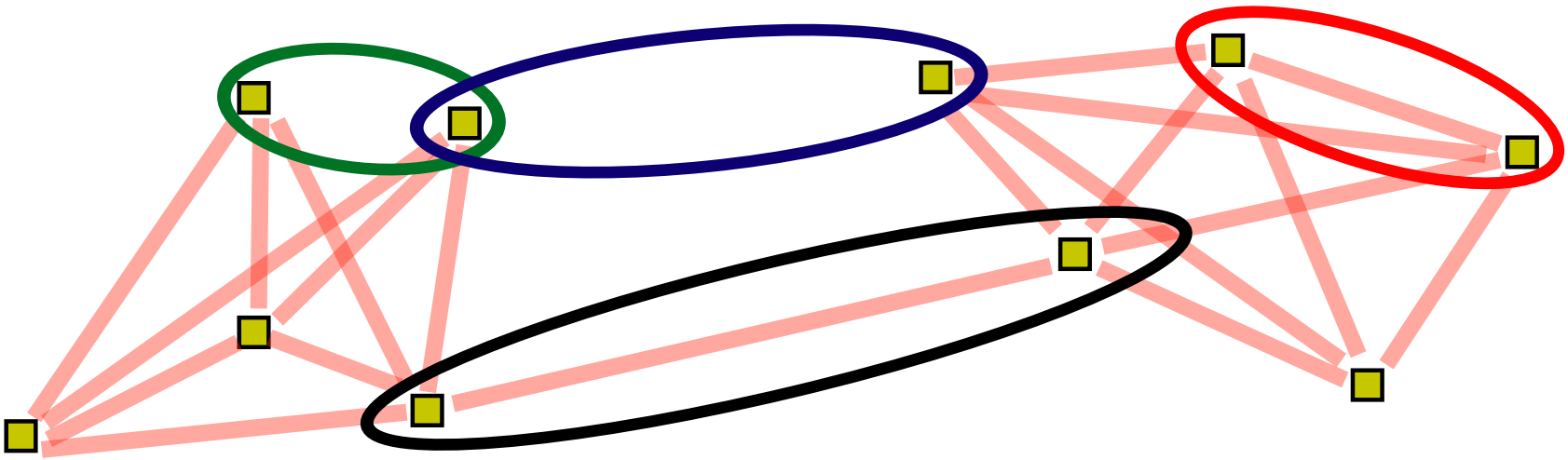
Synonymy	Confluence	
No	weak	Unrelated
No	strong	Potential synonyms
Yes	weak	Shortcut
Yes	strong	Strong synonyms



Signature of confluence

4 types of pairs of vertices:

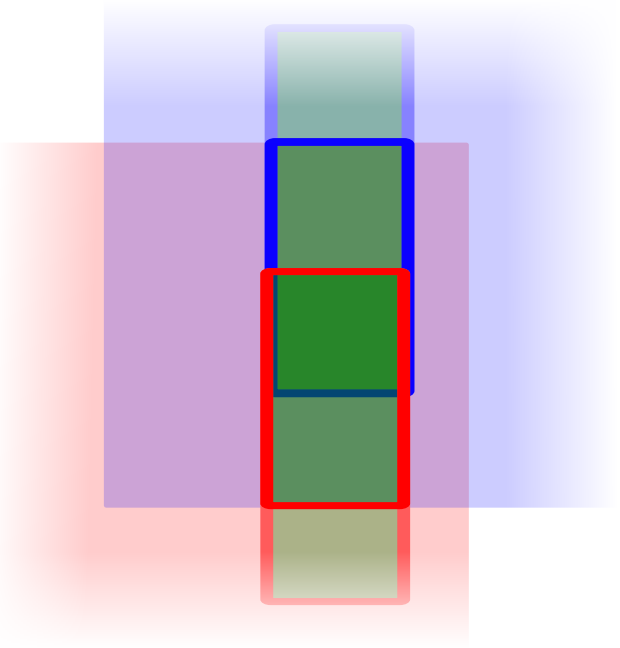
Synonymy	Confluence	
No	weak	Unrelated
No	strong	Potential synonyms
Yes	weak	Shortcut
Yes	strong	Strong synonyms



Self Mediated Agreement by Confluence

Negotiation:

- $(G_1, G_2) \rightarrow (G_1^{+G_2}, G_2^{+G_1})$
- $G_1^{+G_2}$: **add** synonyms of G_2 that are ***potential synonyms*** of G_1
- $G_2^{+G_1}$: **add** synonyms of G_1 that are ***potential synonyms*** of G_2



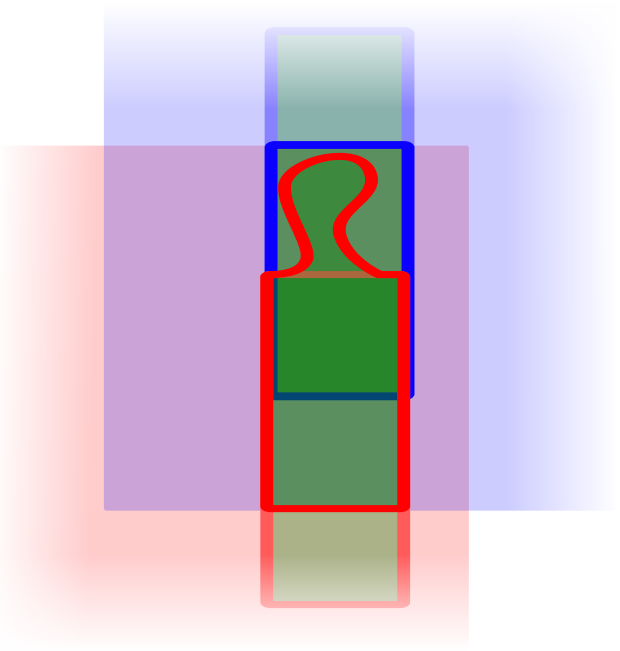
Coarser grain, semantic similarity:

$$SMAC(G_1, G_2) = K_{\updownarrow}(G_1^{+G_2}, G_2^{+G_1})$$

Self Mediated Agreement by Confluence

Negotiation:

- $(G_1, G_2) \rightarrow (G_1^{+G_2}, G_2^{+G_1})$
- $G_1^{+G_2}$: **add** synonyms of G_2 that are ***potential synonyms*** of G_1
- $G_2^{+G_1}$: **add** synonyms of G_1 that are ***potential synonyms*** of G_2



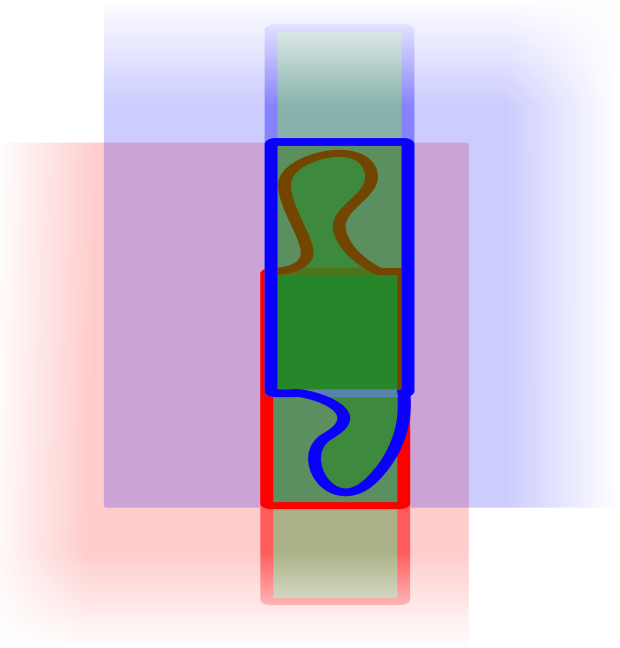
Coarser grain, semantic similarity:

$$SMAC(G_1, G_2) = K_{\updownarrow}(G_1^{+G_2}, G_2^{+G_1})$$

Self Mediated Agreement by Confluence

Negotiation:

- $(G_1, G_2) \rightarrow (G_1^{+G_2}, G_2^{+G_1})$
- $G_1^{+G_2}$: **add** synonyms of G_2 that are ***potential synonyms*** of G_1
- $G_2^{+G_1}$: **add** synonyms of G_1 that are ***potential synonyms*** of G_2



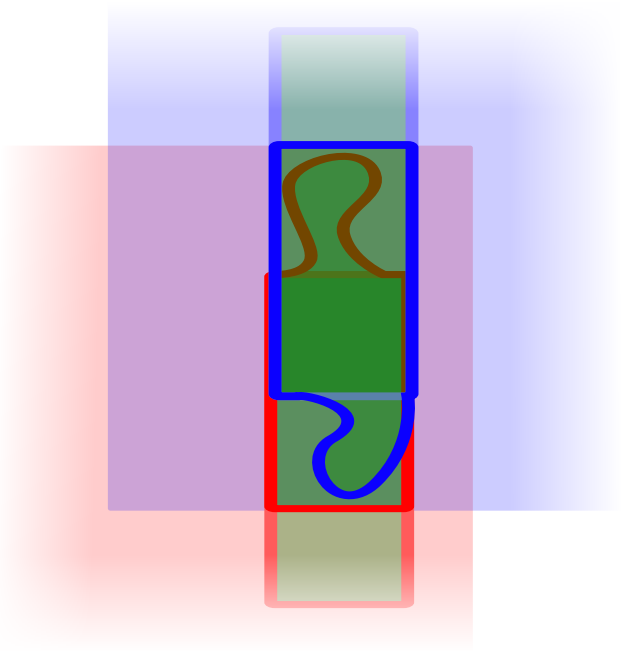
Coarser grain, semantic similarity:

$$SMAC(G_1, G_2) = K_{\updownarrow}(G_1^{+G_2}, G_2^{+G_1})$$

Self Mediated Agreement by Confluence

Negotiation:

- $(G_1, G_2) \rightarrow (G_1^{+G_2}, G_2^{+G_1})$
- $G_1^{+G_2}$: **add** synonyms of G_2 that are ***potential synonyms*** of G_1
- $G_2^{+G_1}$: **add** synonyms of G_1 that are ***potential synonyms*** of G_2



Coarser grain, semantic similarity:

$$SMAC(G_1, G_2) = K_{\updownarrow}(G_1^{+G_2}, G_2^{+G_1})$$

Experimental set up

- Negociation between pairs of graphs (same POS, same language),
- Shortest "possible" walks: $t = 2$,
- Measure Kappa of graphs after negociation,
- Control experiment: negociation of *random graphs* (same edge agreement).

Results

	$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	ori.	0.583	0.309	0.255	0.288	
	acc.	**0.777**	**0.572**	**0.603**	**0.567**	
	ori. r.	0.583	0.309	0.256	0.288	
	acc. r.	0.585	0.313	0.262	0.293	
Ben_V	ori.		0.389	0.276	0.293	
	acc.		**0.657**	**0.689**	**0.636**	
	ori. r.		0.390	0.276	0.294	
	acc. r.		0.392	0.283	0.301	
Ber_V	ori.			0.416	0.538	
	acc.			**0.838**	**0.868**	
	ori. r.			0.417	0.539	
	acc. r.			0.434	0.549	
Lar_V	ori.				0.518	
	acc.				**0.852**	
	ori. r.				0.518	
	acc. r.				0.529	
PWN_V	ori.					0.247
	acc.					**0.540**
	ori. r.					0.247
	acc. r.					0.251

Results

	$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	ori.	0.583	0.309	0.255	0.288	
	acc.	**0.777**	**0.572**	**0.603**	**0.567**	
	ori. r.	0.583	0.309	0.256	0.288	
	acc. r.	0.585	0.313	0.262	0.293	
Ben_V	ori.		0.389	0.276	0.293	
	acc.		**0.657**	**0.689**	**0.636**	
	ori. r.		0.390	0.276	0.294	
	acc. r.		0.392	0.283	0.301	
Ber_V	ori.			0.416	0.538	
	acc.			**0.838**	**0.868**	
	ori. r.			0.417	0.539	
				.434	0.549	
					0.518	
					0.852	
					0.518	
					0.529	
						0.247
PWN_V	**acc.**					**0.540**
	ori. r.					0.247
	acc. r.					0.251

- From medium to strong agreement,
- From weak to medium agreement,
- Control: random networks fail to improve.
- Note: order is not maintained !

Results

	$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	ori.	0.583	0.309	0.255	0.288	
	acc.	**0.777**	**0.572**	**0.603**	**0.567**	
	ori. r.	0.583	0.309	0.256	0.288	
	acc. r.	0.585	0.313	0.262	0.293	
Ben_V	ori.		0.389	0.276	0.293	
	acc.		**0.657**	**0.689**	**0.636**	
	ori. r.		0.390	0.276	0.294	
	acc. r.		0.392	0.283	0.301	
Ber_V	ori.			0.416	0.538	
	acc.			**0.838**	**0.868**	
	ori. r.			0.417	0.539	
				.434	0.549	
					0.518	
					0.852	
					0.518	
					0.529	
PWN_V	**acc.**					0.247
	ori. r.					**0.540**
	acc. r.					0.247
						0.251

- From medium to strong agreement,
- From weak to medium agreement,
- Control: random networks fail to improve.
- Note: order is not maintained !

Results

	$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	ori.	0.583	0.309	0.255	0.288	
	acc.	**0.777**	**0.572**	**0.603**	**0.567**	
	ori. r.	0.583	0.309	0.256	0.288	
	acc. r.	0.585	0.313	0.262	0.293	
Ben_V	ori.		0.389	0.276	0.293	
	acc.		**0.657**	**0.689**	**0.636**	
	ori. r.		0.390	0.276	0.294	
	acc. r.		0.392	0.283	0.301	
Ber_V	ori.			0.416	0.538	
	acc.			**0.838**	**0.868**	
	ori. r.			0.417	0.539	
				.434	0.549	
					0.518	
					0.852	
					0.518	
					0.529	
PWN_V	**acc.**					0.247
	ori. r.					**0.540**
	acc. r.					0.247
						0.251

- From medium to strong agreement,
- From weak to medium agreement,
- Control: random networks fail to improve.
- Note: order is not maintained !

Results

	$K_{\updownarrow}$	Ben_V	Ber_V	Lar_V	Rob_V	Wik_V
Bai_V	ori.	0.583	0.309	0.255	0.288	
	acc.	**0.777**	**0.572**	**0.603**	**0.567**	
	ori. r.	0.583	0.309	0.256	0.288	
	acc. r.	0.585	0.313	0.262	0.293	
Ben_V	ori.		0.389	0.276	0.293	
	acc.		**0.657**	**0.689**	**0.636**	
	ori. r.		0.390	0.276	0.294	
	acc. r.		0.392	0.283	0.301	
Ber_V	ori.			0.416	0.538	
	acc.			**0.838**	**0.868**	
	ori. r.			0.417	0.539	
				.434	0.549	
					0.518	
					0.852	
					0.518	
					0.529	
PWN_V	**acc.**					0.247
	ori. r.					**0.540**
	acc. r.					0.247
						0.251

- From medium to strong agreement,
- From weak to medium agreement,
- Control: random networks fail to improve.
- Note: order is not maintained !

Conclusions and Perspectives

Conclusions:

- Global agreement on semantic structures of resource
  vs. pair by pair synonymy variability
- Beyond synonymy:
  confluence as a global, gradient measure of word's semantic similarity
- Confluence based on synonymy resources
  but relatively independent from synonymy particulars

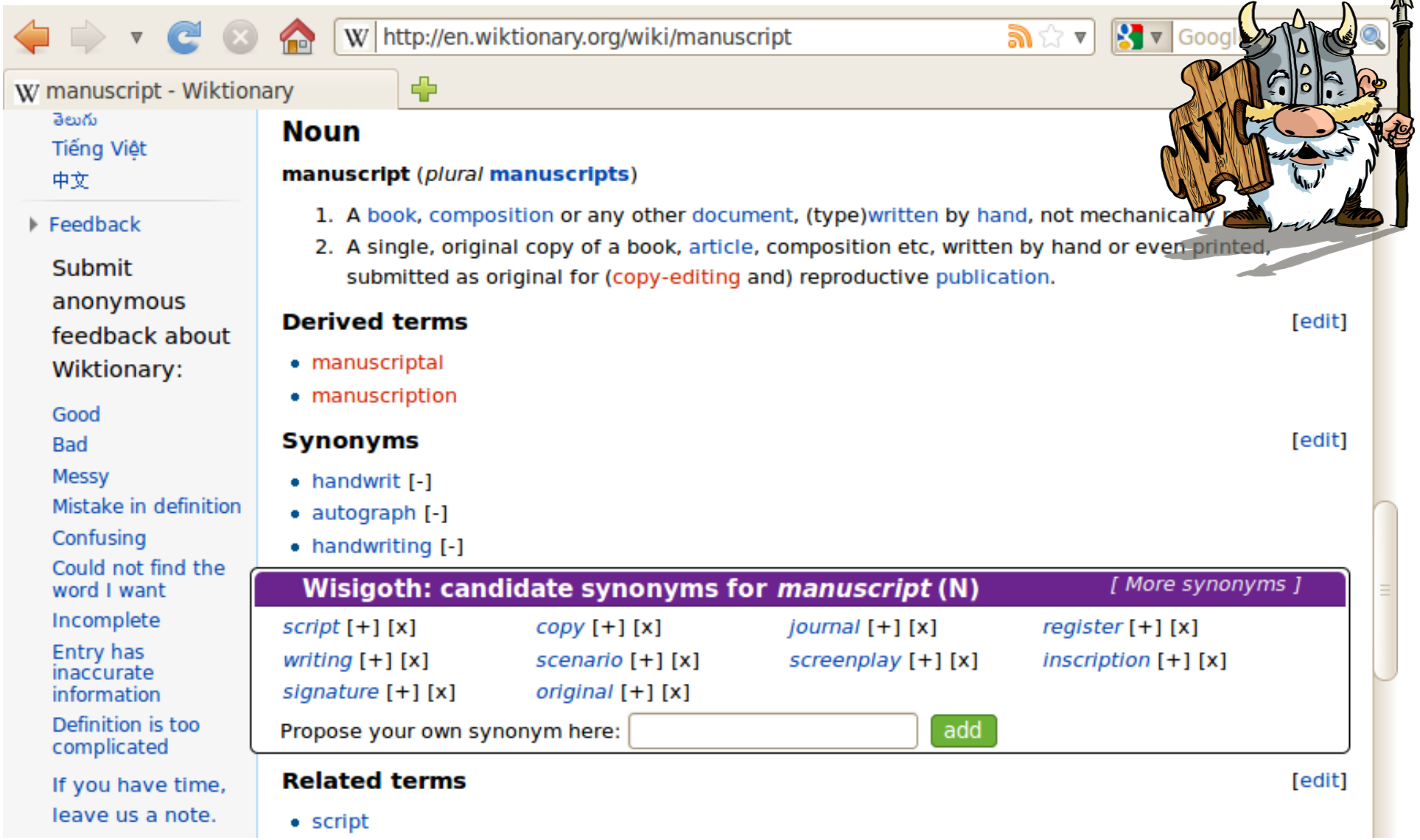
Perspectives:

- Vary t (random walk length)
  and k (ratio of probability to limit for strong/weak confluence)
- Balanced negociation: not only add proposed potential synonyms,
  but also remove declined shortcuts

Applications

- **Merging** resources:
  between union and intersection, filtering shortcuts,
- **Evaluating** resources:
  Semantic agreement of new resources with (sets of) established resources,
- **Bilingual** semantic comparison:
  via translation bigraph, SMAC of graphs of different languages,
- Semi-automatic **enrichment** of Wiktionary by confluence.

Application to semi-automatic enrichment of Wiktionary



WISIGOTH Project : `http://redac.univ-tlse2.fr/wisigoth/`

谢谢