Dynamiques des réseaux sociaux en ligne recommandations et interactions

Stéphane Raux

LIAFA – Linkfluence

Séminaire *Complex Networks*

Motivations

- Les plateformes de réseaux sociaux
  - Popularité de la notion de « réseau » auprès du grand public
  - Ces sites s'appuient sur les interactions entre les utilisateurs
- Changements dans la manière dont nous utilisons le web
  - Recherche et partage d'information
  - Communication
  - Gestion de son image en ligne
- Terrains d'études très riches pour les chercheurs
  - Volumes très importants de données disponibles
  - Données souvent très détaillées (timestamp, API)
  - Accessibilité des sujets d'étude

Linkfluence

- PME spécialisée dans la collecte et l'analyse des prises de parole sur le web
  - Institut d'études
    - Etudes d'images, mesure d'efficacité de campagnes de communication
    - Ces études peuvent être complémentaires avec des études traditionnelles
  - Logiciel de veille et d'engagement proposé en *SaaS*
  - Technologies de captation, d'enrichissement et d'indexation de données

Plan

Les données des commentaires de Flickr

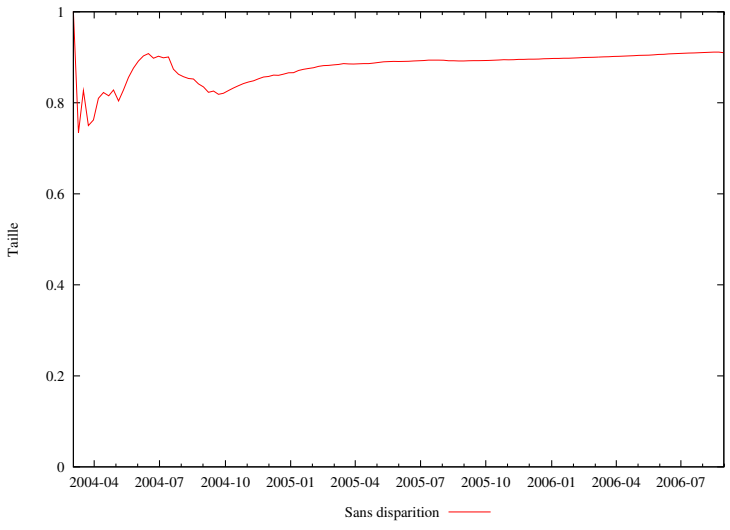
- Projet Autograph (Liafa, Orange Labs, INRIA ...)
- Extraction en août-septembre 2006 par Orange Labs
- 156 840 996 photographies (publiques), 4 788 438 utilisateurs
- Le graphe des commentaires :
  39 594 157 commentaires, 910 454 utilisateurs
- De mars 2004 à fin août 2006

Comment construire le graphe ?

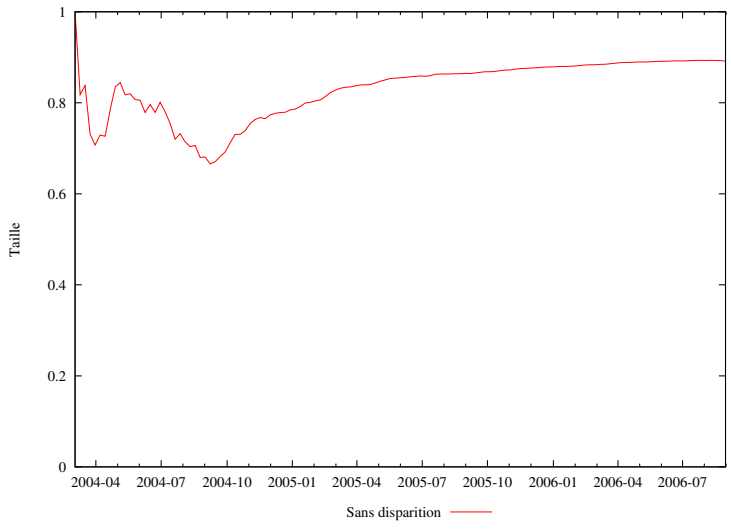
- Ajout de tous les commentaires au fil du temps
- Réduction aux liens réciproques :

	Sommets	Commentaires
Graphe complet	910 454	39 594 157
Liens réciproques	259 395	25 818 794
	28, 4%	65, 2%

Taille relative de la composante connexe principale



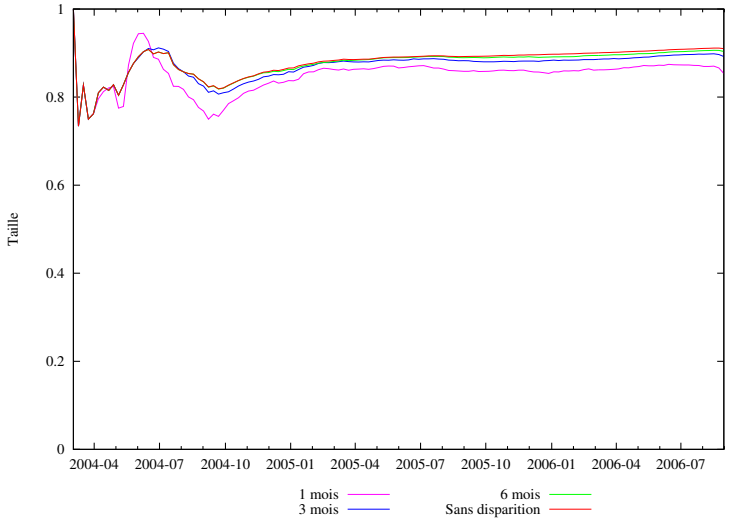
Graphe complet



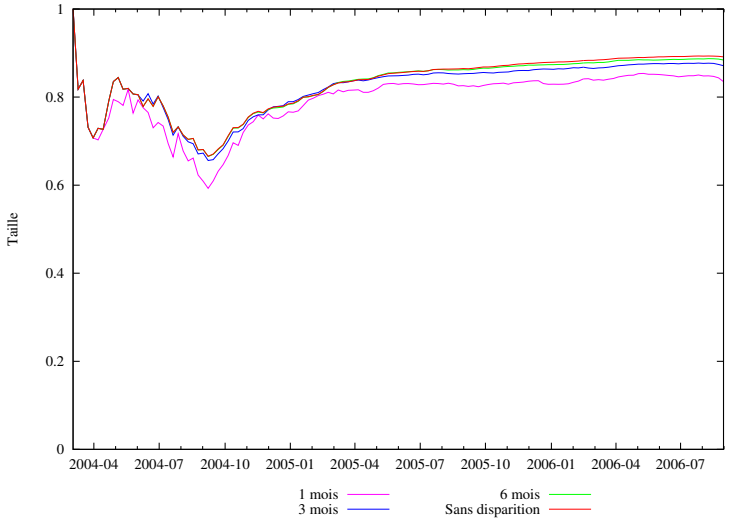
Graphe des liens réciproques

- L'évolution de la structure globale diffère peu dans les deux cas

Taille relative de la composante connexe principale



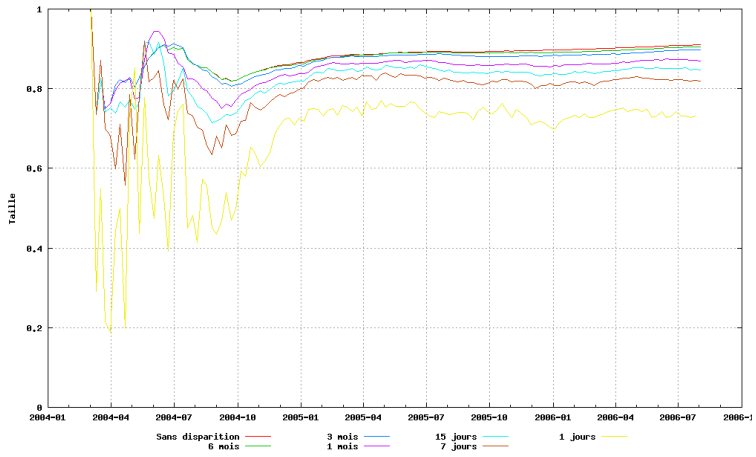
Graphe complet



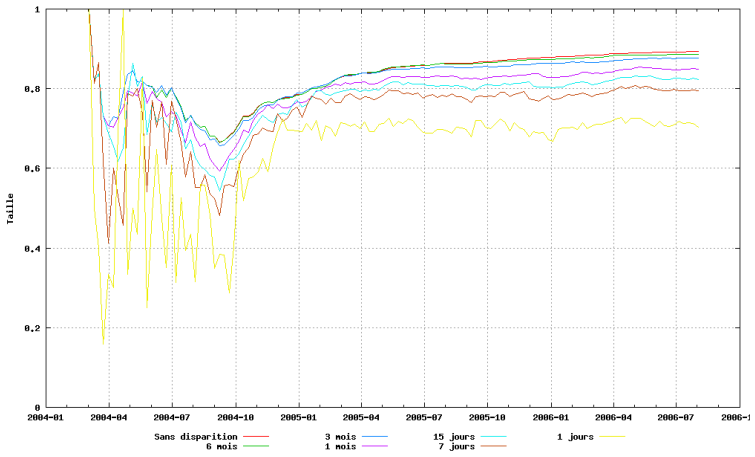
Graphe des liens réciproques

- Prise en compte de la « durée de vie » des relations
  - Peu de différences pour une durée de 1, 3 ou 6 mois
  - Les différences sont sensibles en dessous de 7 jours

Taille relative de la composante connexe principale



Graphe complet



Graphe des liens réciproques

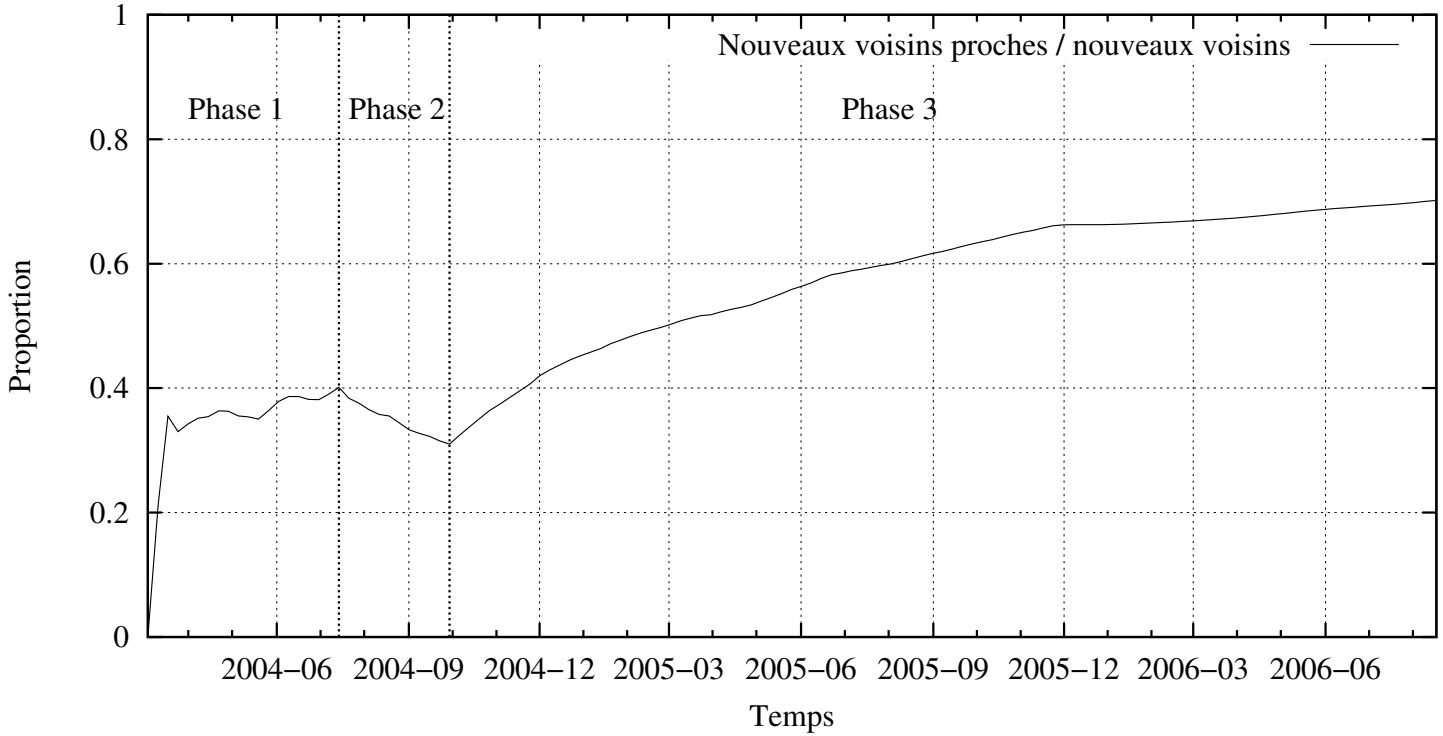
- Prise en compte de la « durée de vie » des relations
  - Peu de différences pour une durée de 1, 3 ou 6 mois
  - Les différences sont sensibles en dessous de 7 jours

Répartition des commentaires en fonction du type de contact

Répétitions	Nouveaux voisins		Total
	Proches	Lointains	
Distance $= 1$	Distance $= 2$	Distance ≥ 3	
29 946 674	6 781 686	2 865 797	39 594 157
$75,6\%$	$17,2\%$	$7,2\%$	100%

- 93% des commentaires sont des répétitions ou se font entre voisins proches
- Les voisins proches représentent 70% des nouveaux voisins

Évolution de la proportion de voisins proches



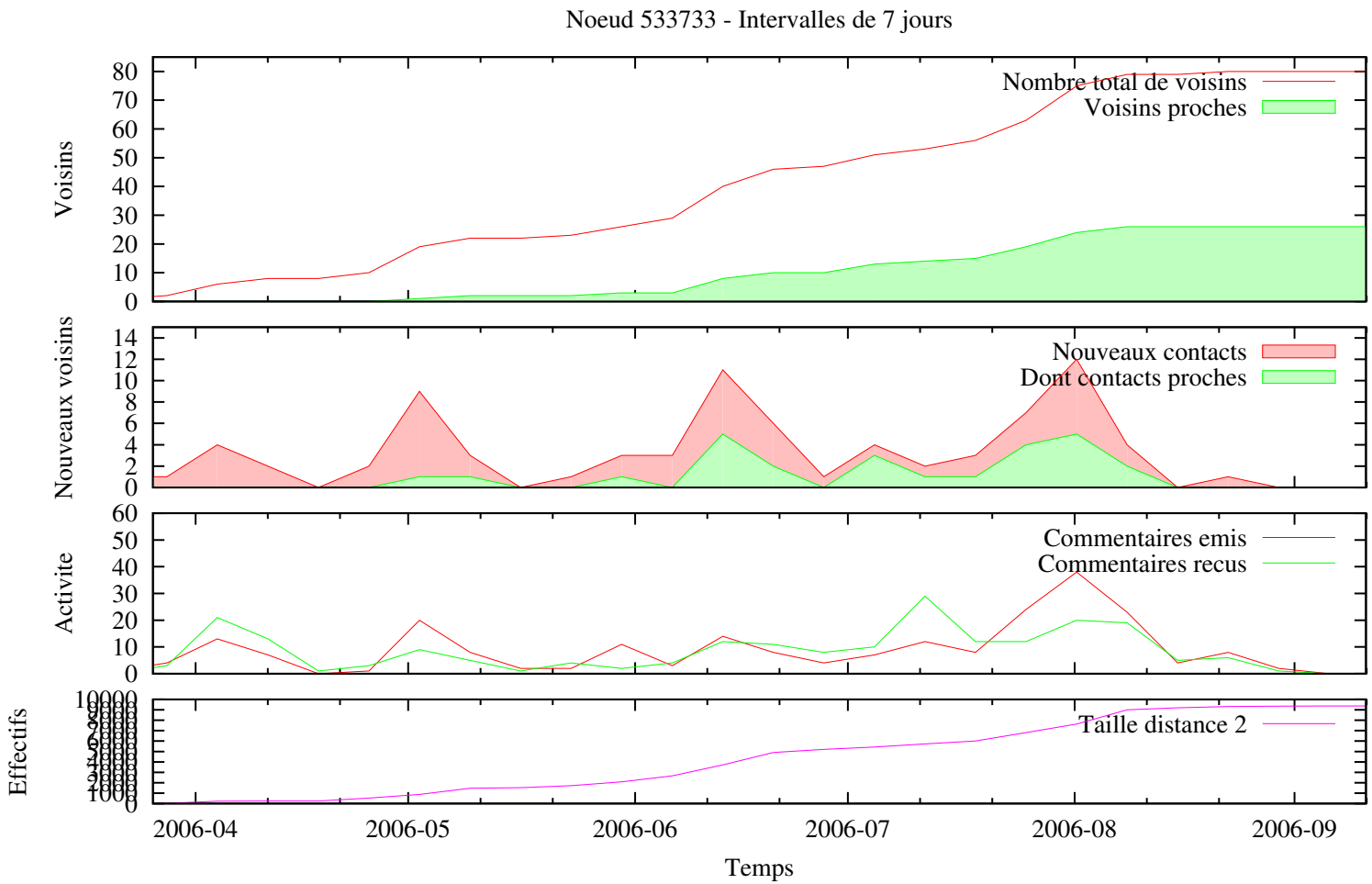
- Evolution en trois grandes étapes
- Résultats similaires obtenus par Kumar et Tomkins (2006) sur Twitter

Suivi individuel de deux individus

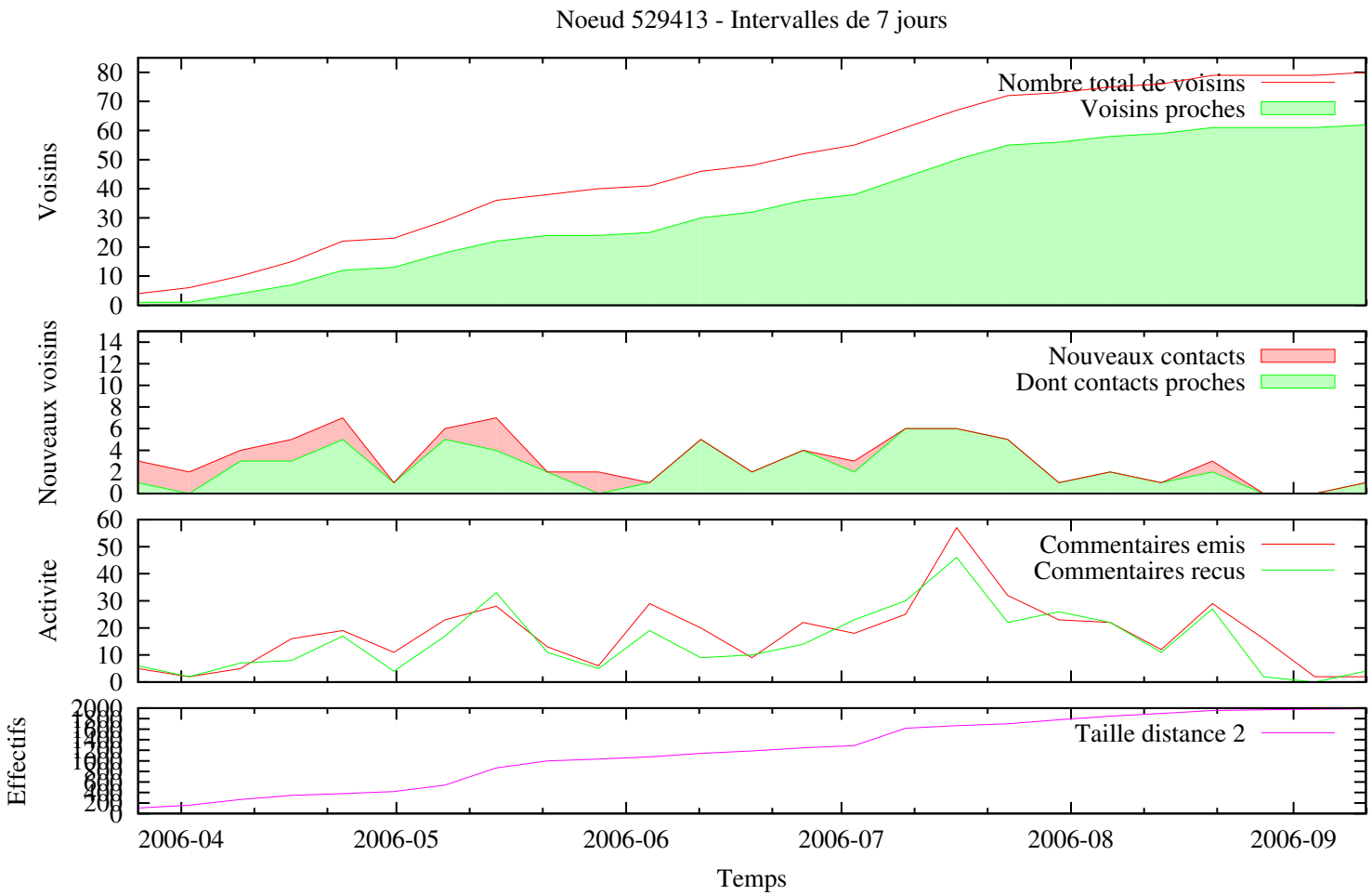
On choisit deux individus :

- Degré 80 dans le graphe des commentaires réciproques
- Même période d'activité, à partir de fin mars 2006
- L'un privilégie les voisins proches, l'autre les voisins lointains

Évolution de l'entourage de l'individu *A*, qui privilégie les contacts lointains



Évolution de l'entourage de l'individu *B*, qui privilégie les contacts proches

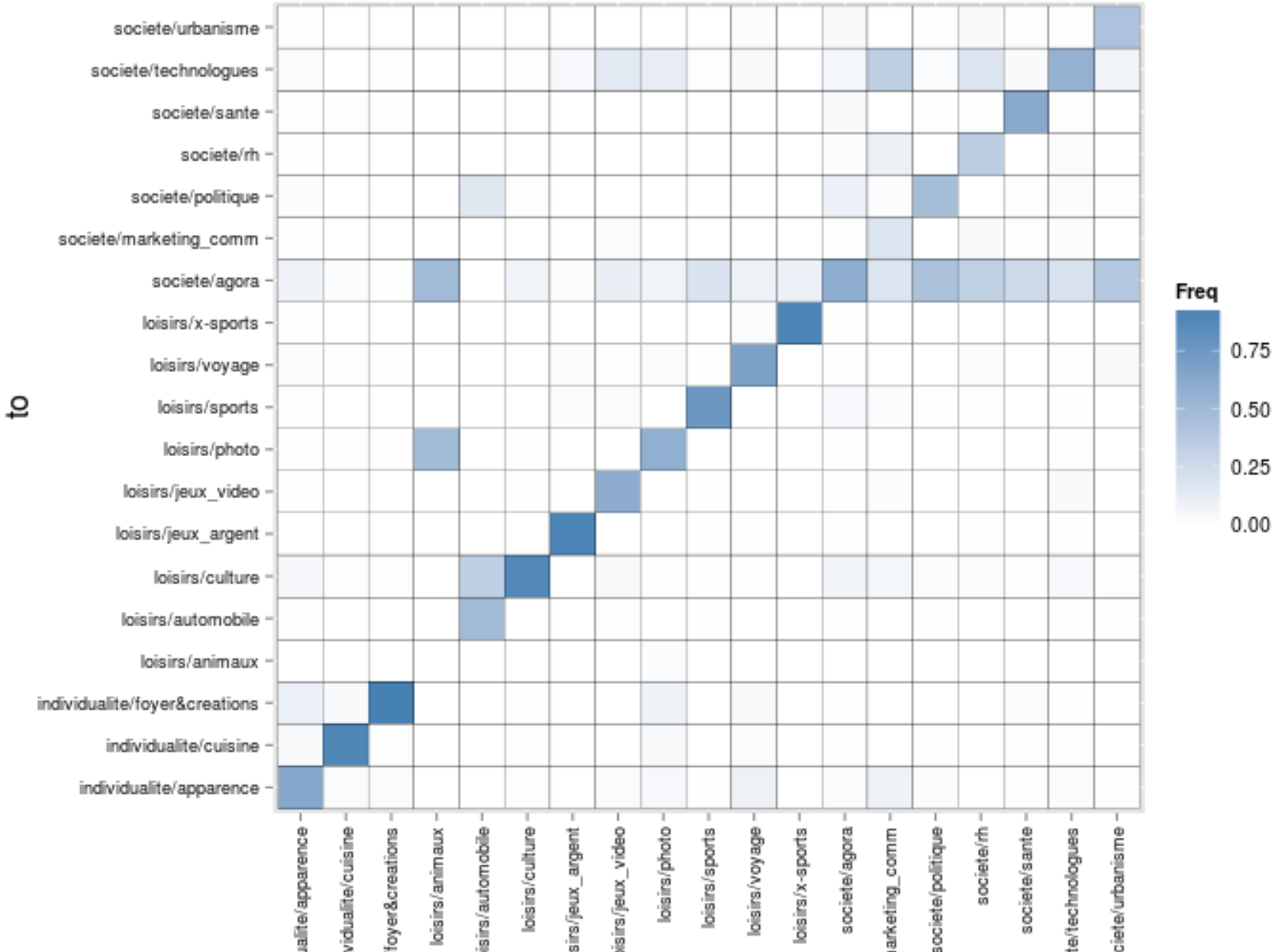


Plan

Segmentation du web en communautés

- Communautés définies selon des critères topologiques et sémantiques
- Elles sont maintenues par des documentalistes, pour plus de six pays
- Hiérarchie sur trois niveaux (continent, territoires, communautés)

Répartition des citations entre communautés

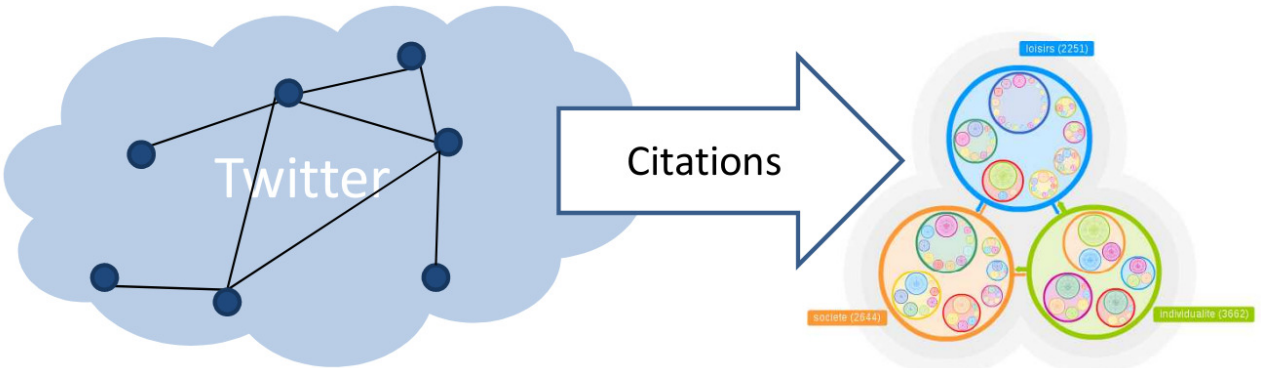


Objectifs

- Constituer un multi-réseau blogs – Twitter
- On conserve les mêmes principes que pour l'écosphère
  - Échantillonage
  - Catégorisation
  - Analyse différentielle

Méthode

- Notre Démarche :
  - On recherche tous les *tweets* qui contiennent une URL vers l'écosphère
  - On sélectionne parmi les auteurs de ces *tweets* un ensemble d'utilisateurs dont on va suivre tous les messages
- Contraintes :
  - Très grand volume de publications
  - Bruit important lié au caractère international de l'audience de certains sites



Méthode naïve : Fixer un seuil

- On fixe un seuil N
- On ne retient que les utilisateurs qui ont cité au moins N sites dans l'écosphère : $|S(u)| \geq N$
- On écarte ainsi les utilisateurs étrangers (citations ponctuelles) et les utilisateurs qui ne font que de l'auto-promotion

Évaluer la « confiance »qu'on accorde à chaque site

- On calcule pour chaque utilisateur un score de *h-index* : un utilisateur a un indice h s'il a au moins cité h sites différents h fois chacun.
- On calcule pour chaque site $C(s)$, la proportion d'utilisateurs qui ont un *h-index* supérieur ou égal à deux

Répercussion de ces scores sur les utilisateurs

- Le score de confiance $\sigma(u)$ de chaque utilisateur u correspond à la somme des scores de confiance des sites qu'il a cités

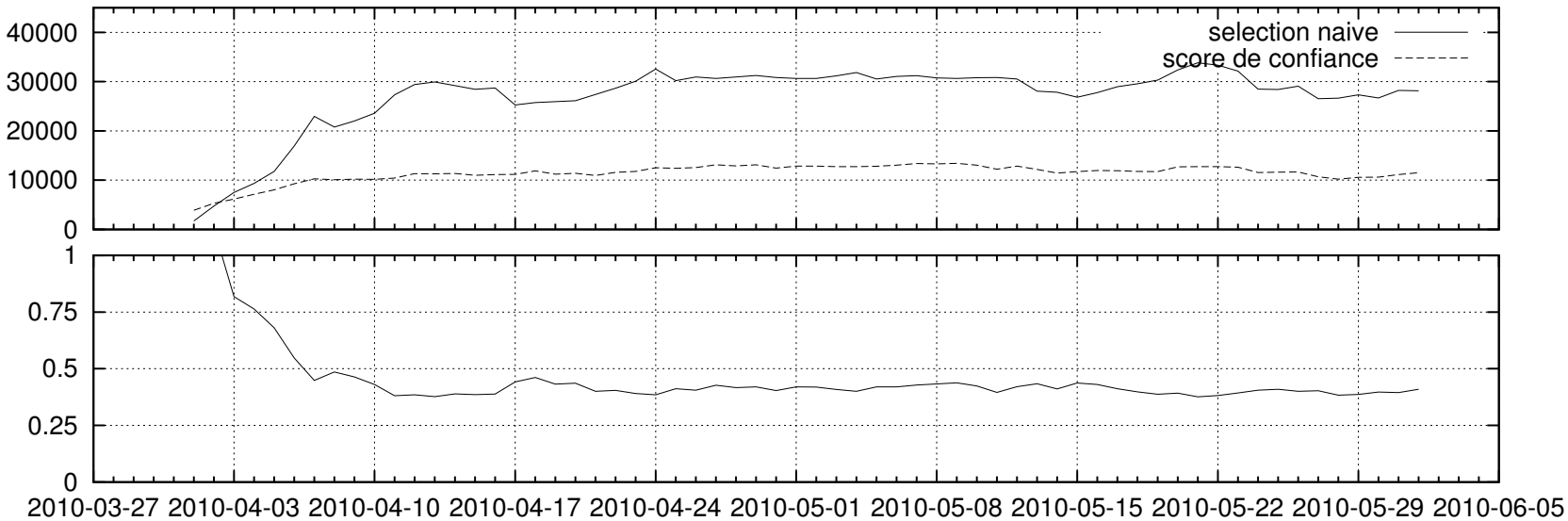
$$\sigma(u) = \sum_{s \in S(u)} C(s)$$

- On calcule la médiane des scores de confiances non-nuls des sites
- On ne retient que les utilisateurs dont le score de confiance est supérieur à cette médiane
  - Utilisateurs qui ont cité un site en qui on a confiance
  - Utilisateurs qui ont cités plusieurs sites moins « sûrs »

Simulations de la sélection

- Période du 1er avril 2010 au 1er juin 2010
- Une mesure par jour, sur l'intervalle des 7 derniers jours
- Pour chaque utilisateur, on retient :
  - Le nombre de *tweets*
  - Le nombre de sites cités
  - Le nombre d'URL citées
- On compare ces mesures pour l'ensemble des utilisateurs du corpus et pour les utilisateurs sélectionnés

Comparaison du nombre total d'utilisateurs sélectionnés



- Les deux méthodes donnent des résultats similaires en termes de proportions d'utilisateurs sélectionnés chaque jour
- Le calcul des scores de confiance permet de diversifier davantage les sites et les URL cités, en sélectionnant un moins grand nombre d'utilisateurs à l'échelle de la semaine

Comparaison du nombre total d'utilisateurs sélectionnés

- Les deux modes de sélection conservent bien les hiérarchies de citations
- La mesure serait plus pertinente si on comparait les coefficients à ceux obtenus par une sélection aléatoire

Algorithme de reconnaissance du langage

- L'écosphère et le Twitter français se caractérisent par un emploi majoritaire du français
- Algorithme de reconnaissance du français et de l'anglais (comparaison de préfixes, suffixes et trigrammes)
- Préformatage des *tweets* :
  - On retire les URL, les mentions et les hashtags
  - On ne traite que les messages qui contiennent au moins 4 mots après formatage

Distribution *tweets* en fonction des langages détectés

	Français	Inconnu	Anglais	Total
Corpus	572 205 62%	296 527 32%	52 264 6%	920 996
Sélection	460 768 69%	188 401 28%	21 229 3%	670 398

- La sélection par les scores de confiance accentue la prépondérance des *tweets* en français
- La proportion de messages en anglais chute de moitié

Plan

Corrélations entre indicateurs d'activité sur Twitter

	tweets	url	rt	is_rt	at
url	0.82				
rt	0.67	0.54			
is_rt	0.22	0.15	0.12		
at	0.69	0.21	0.37	0.19	
is_at	0.54	0.11	0.22	0.41	0.83

- Forte corrélation (0.54) entre le nombre d'urls citées et le nombre de *retweets*
- Forte corrélation (0.83) entre mention envoyées et reçues
- Corrélation moyenne entre *retweets* et mentions reçues (0.41)

Typologie en fonction des indicateurs d'activité

groupe	quantile	tweets	url	rt	is_rt	at	is_at
1	mediane	341	128	119	71	157	165
	90%	695	240	217	150	365	395
2	mediane	66	22	14	15	32	35
	90%	120	42	26	34	67	73
3	mediane	48	35	25	4	8	3
	90%	88	67	50	9	17	6
4	mediane	160	155	1	9	7	4
	90%	338	334	3	21	11	10
5	mediane	9	5	2	1	2	2
	90%	17	11	6	3	6	5

- Utilisation de la méthodes des *k-means* sur tweets, url et rt
- cinq groupes : 1) « starts », 2) « experts », 3) « veilleurs », 4) « robots », 5) « inactifs »

Contributions

- Mesure de grands réseaux, à une échelle globale
  - Analyse des commentaires de Flickr
  - Echantillonage des utilisateurs de Twitter
- Analyse de réseaux à une échelle locale
  - Evolution des voisinages sur Flickr
  - Proposition de communautés pour corpus Linkfluence
  - Typologie des utilisateurs de Twitter
  - Application Algopol
- Enrichissement des données collectées sur Twitter
  - Détection de sujets à partir des citations de liens
  - Reconstitution de cascades de diffusion d'urls