Une algorithmique des graphes de terrain

Maximilien Danisch

Post-Doc à TélécomParisTech

Les graphes de terrain

Définition

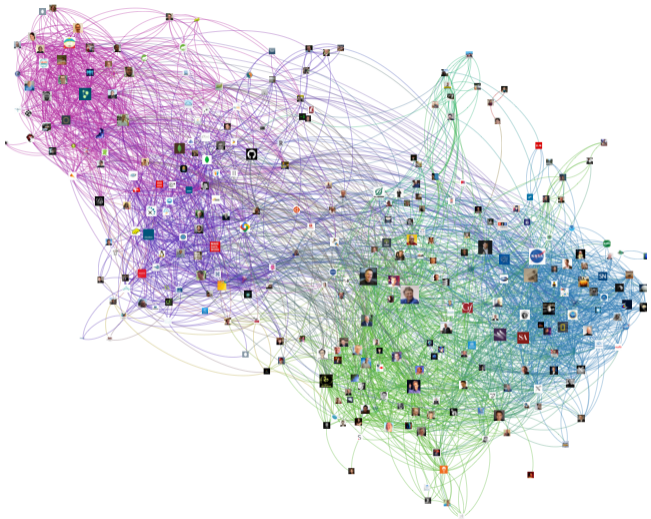
Graphe. Un ensemble de nœuds connectés par des liens.

graphes	nœuds	liens
Facebook	profils	amitiés
Internet	ordinateurs	connections
Web	pages web	hyperliens
Cerveaux	neurones	synapses
Encore plus général		

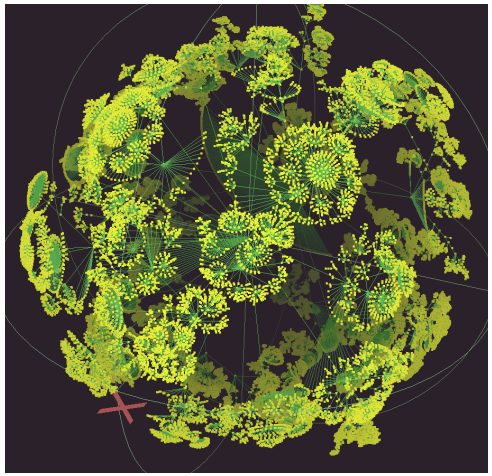
Propriétés empiriques

- Très grands
- Peu denses
- Triangles
- Degrés hétérogènes

Twitter



Internet



- **Besoin d'algorithmes efficaces sur les graphes de terrain.**

Pourquoi une algorithmique des graphes de terrain ?

Structure particulière ⇒ Algorithmique particulière

- Trouver une clique maximum : NP-difficile mais "facile".
  Algorithme glouton + Branch & Bound. Rossi et al. WWW2014.
- Complexité polynomiale : pas forcément bien.
  Longueur de tous les plus courts chemins $\Theta(n^3)$. Floyd-Warshall.
- Complexité en pratique souvent bien meilleure que le pire cas.
  Convergence de l'algorithme de Louvain. Blondel et al. JSTAT2008.

Problèmes avec l'algorithmique de graphes "classique"

- Analyse en complexité dans le pire cas ou en moyenne pas adaptée.
- Algorithmes pour des classes de graphes pas utilisables tels quels.

Mon projet

Mieux comprendre la structure des graphes de terrain et en tirer parti pour proposer de meilleurs algorithmes.

Axes de recherches

1. **Caractériser finement les graphes de terrain**
   - Les propriétés classiques sont trop générales
   - Identifier plusieurs types de graphes de terrain
   - Complexité paramétrée en fonction des propriétés
2. **Adaptation des techniques algorithmiques classiques**
   - Optimisation Convexe (WWW2017)
   - Branch & Bound (Workshop ICDM2016)
3. **Passage à l'échelle : compression + outils big data**
   - Compresser en autorisant des opérations basiques
   - Parallélisation des algorithmes : Spark et Hadoop

Exemple 1. Lister toutes les k-cliques

Pourquoi lister toutes les k-cliques ?

Des travaux récents appellent à une telle subroutine :

- k-clique densest subgraph *Charalampos WWW2015,*
- percolated k-clique communities *Palla et all. Nature2005.*

Contributions principales

- Meilleur algorithme pour lister les k-cliques
- Lister les k-cliques passe mieux à l'échelle que ce que pense la communauté scientifique *Jain et Seshadhri WWW2017.*

Exemple 1. Lister toutes les k-cliques

Algorithme : performance en théorie

- c = core value (ou dégénérescence) du graphe
- **Théorème.** Complexité en $O(m \cdot (\frac{c}{3})^{k-2})$
- Meilleure complexité pour les graphes où $c \ll n$

Algorithme : performance en pratique

- 1 ODG plus rapide que l'état de l'art et parallélisme optimal
- 10-cliques dans Friendster (1.8G liens) : 487 090 833 092 739
- 5-cliques dans Twitter09 (1.6G liens) : 3 388 795 307 518 264

k-cliques streaming

- **Problème.** Trop de k-cliques pour les mettre en mémoire
- **Solution.** Faire du calcul à la volée

Exemple 2. Density-Friendly Decomposition (WWW2017)

Pourquoi s'intéresser à la density-friendly decomposition ?

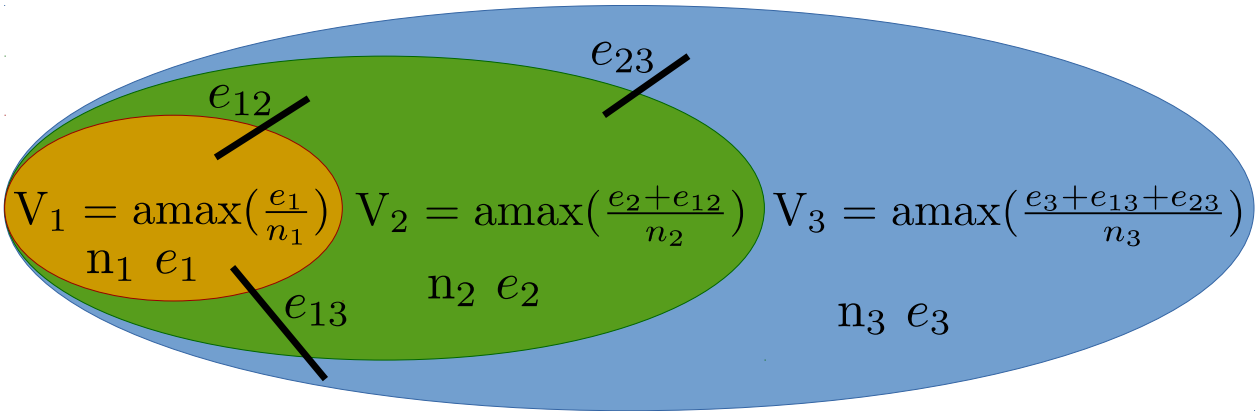
- La density-friendly decomposition (Tatti et Gionis WWW2015) est intéressante car elle fusionne les deux concepts :
  1. core decomposition et
  2. dense subgraph.
- Problème : elle est calculée par une succession de maxflows et ne passe pas à l'échelle.

Contributions principales

- Faire passer à l'échelle la density-friendly decomposition.
- Lien original avec l'optimisation convexe.

Exemple 2. Density-Friendly Decomposition (WWW2017)

Définition. Collection d'ensembles disjoints de nœuds V_i :



Optimisation quadratique

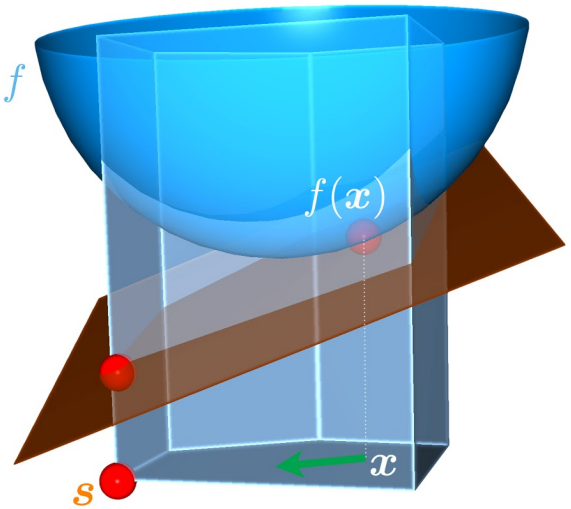
$$\min \quad \sum_{u \in V} r_u{}^2$$

$$\text{s.t.} \quad \forall u \in V,\ r_u = \sum_{e : u \in e} \alpha_u^e$$

$$\forall e \in E,\ \sum_{u \in e} \alpha_u^e = w_e$$

$$\forall u \in e \in E,\ \alpha_u^e \geq 0$$

Algorithme de Frank-Wolfe



Performance en pratique

- 10M liens en quelques secondes
- 2G liens en quelques heures
- 25G liens approx. en quelques heures

Meilleure que le pire cas

Exemple 3. Autour des graphes parfaits

Graphe parfait

- Définition : pas de trou de taille impaire et ≥ 5 dans le graphe et dans son complément.
- Certains problèmes NP-difficiles (e.g. maximum clique) sont résolubles en temps polynomial sur un graphe parfait.

Idée

Modifier ces algorithmes pour qu'ils donnent, en plus, une solution sur des graphes non-parfaits.

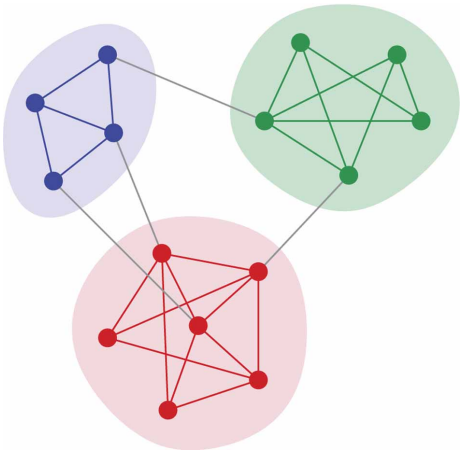
Problème et solution

- **Problème.** Besoin SDP efficaces (fonction thêta de Lovász), les méthodes classiques sont trop lentes (MOSEK).
- **Solution.** Frank-Wolfe et Sparse-SDP. *Jaggi ICML2013.*

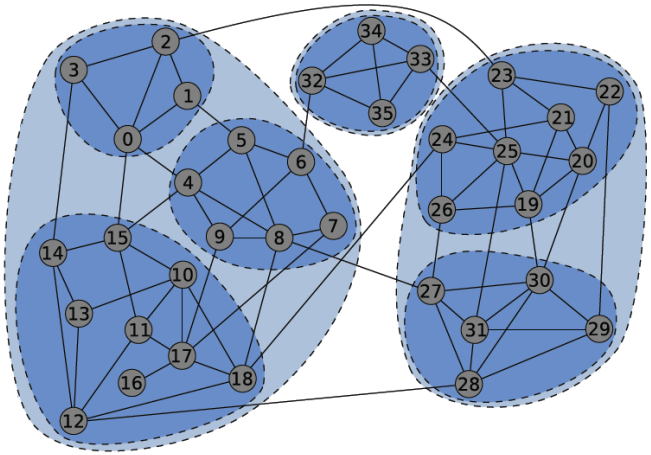
Thèse. Communautés locales et mesures de proximité

Vision des communautés avant ma thèse :

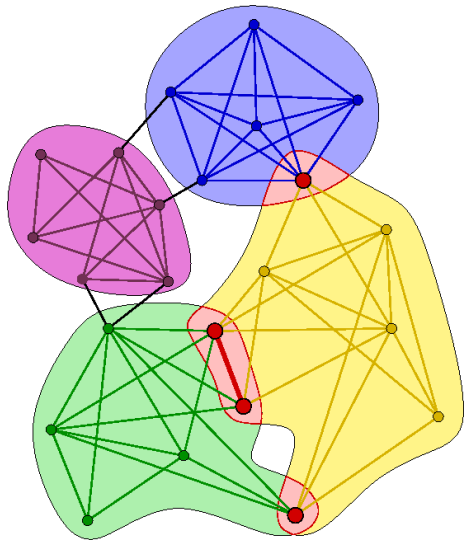
partition



hierarchiques

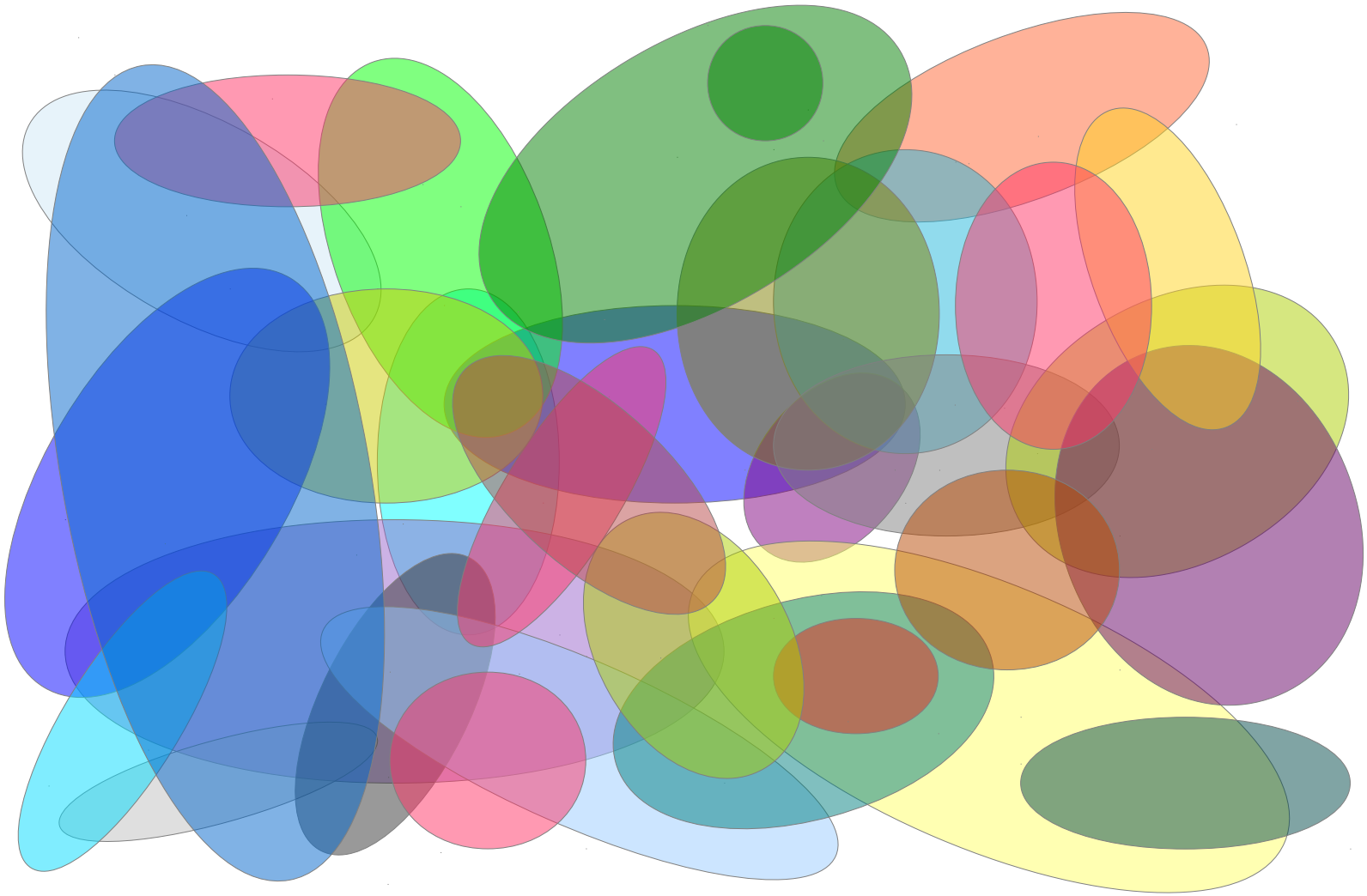


recouvrantes

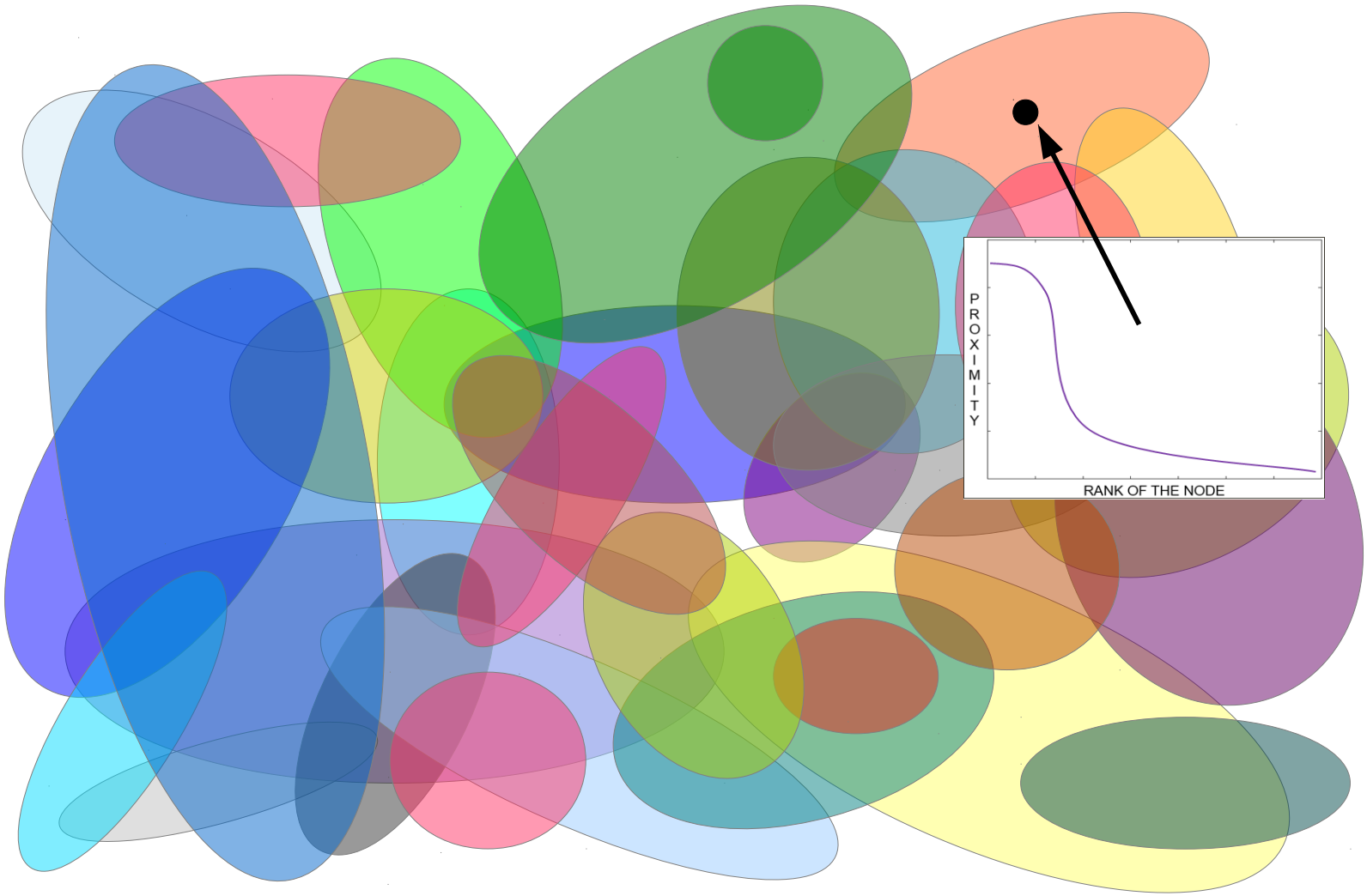


Approche classique pour trouver des communautés :
optimisation d'une fonction de qualité ad hoc

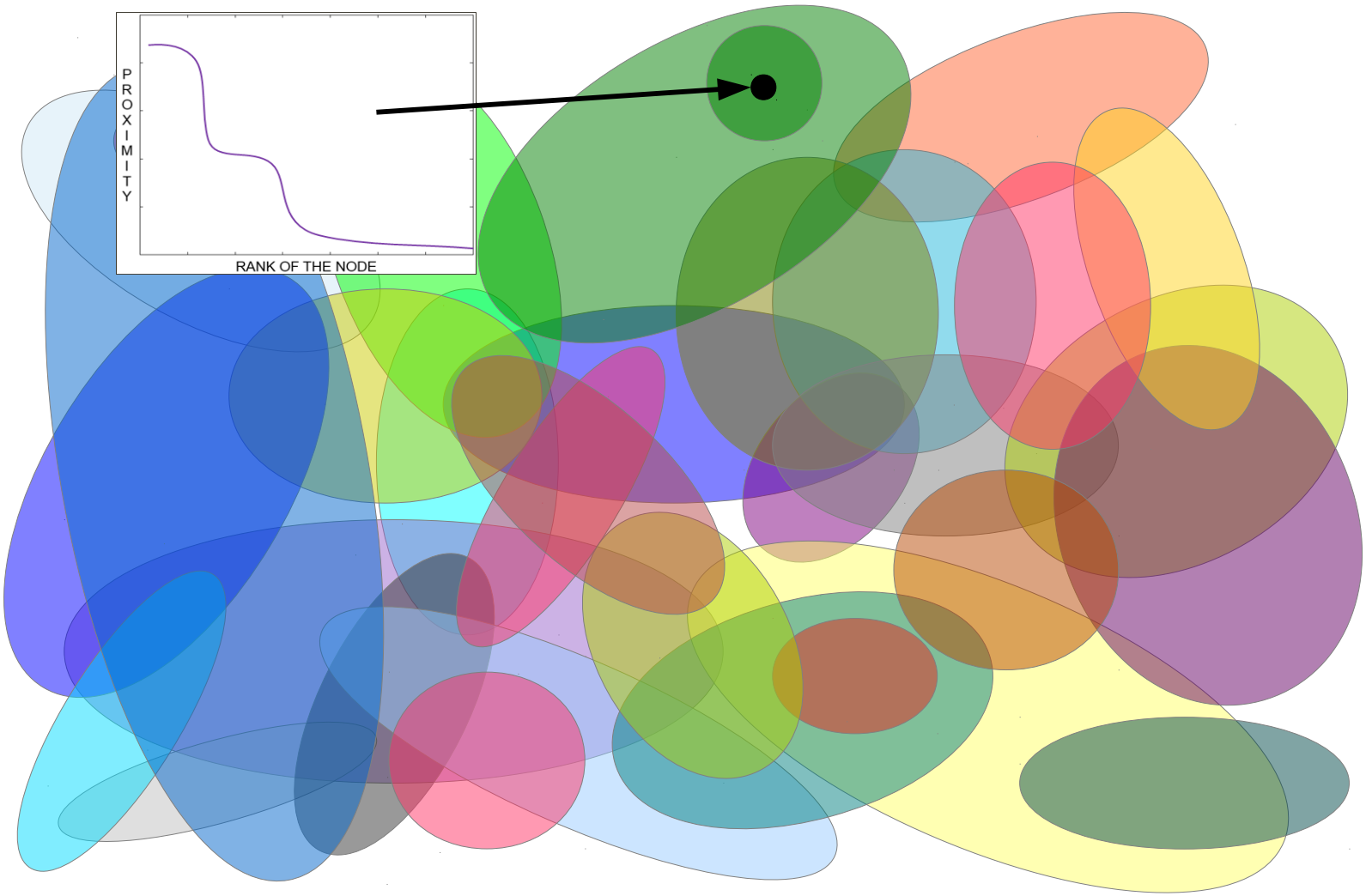
Thèse. Communautés locales et mesures de proximité



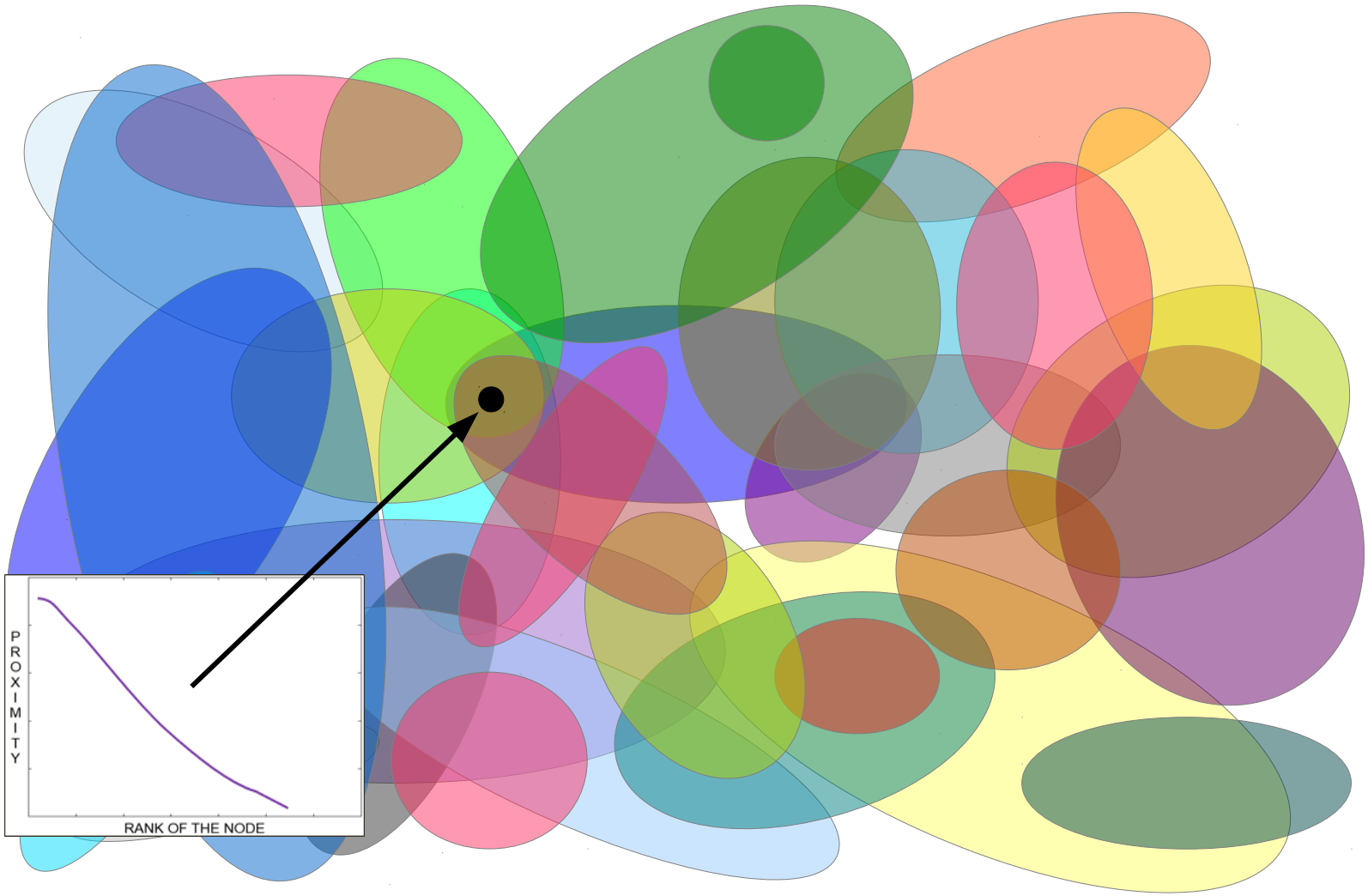
Thèse. Communautés locales et mesures de proximité



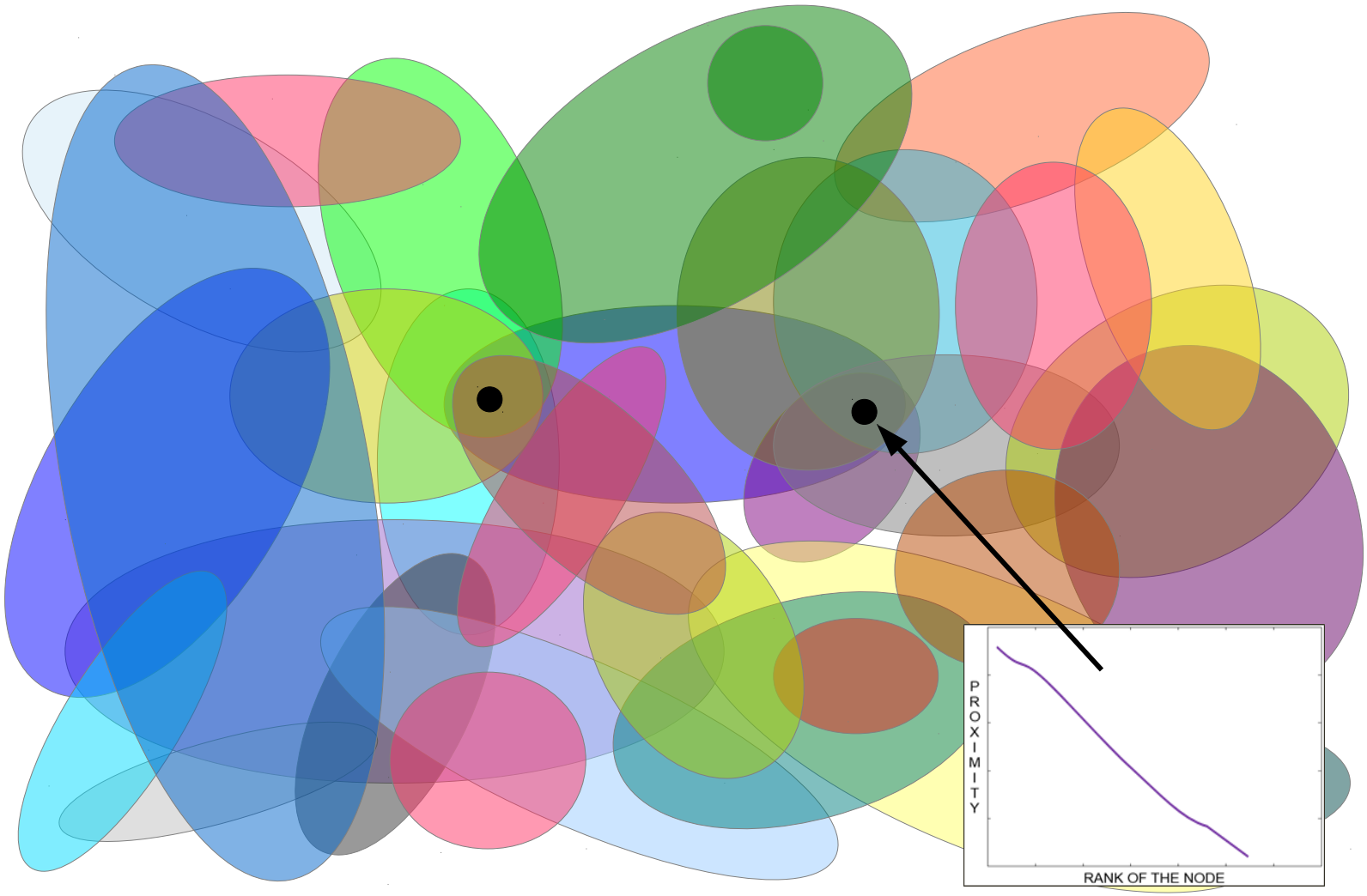
Thèse. Communautés locales et mesures de proximité



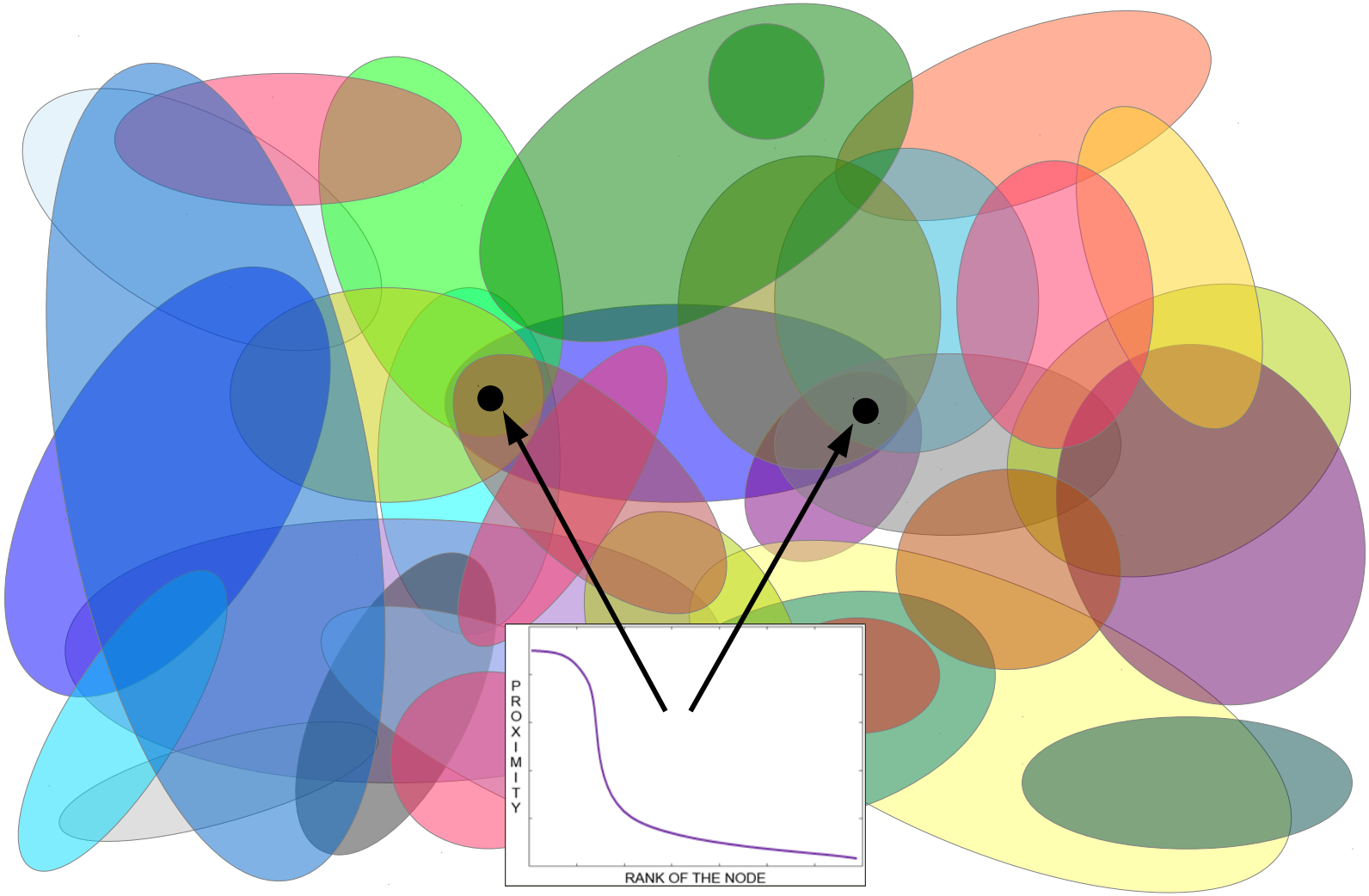
Thèse. Communautés locales et mesures de proximité



Thèse. Communautés locales et mesures de proximité



Thèse. Communautés locales et mesures de proximité



Thèse. Communautés locales et mesures de proximité

Validation de l'approche sur

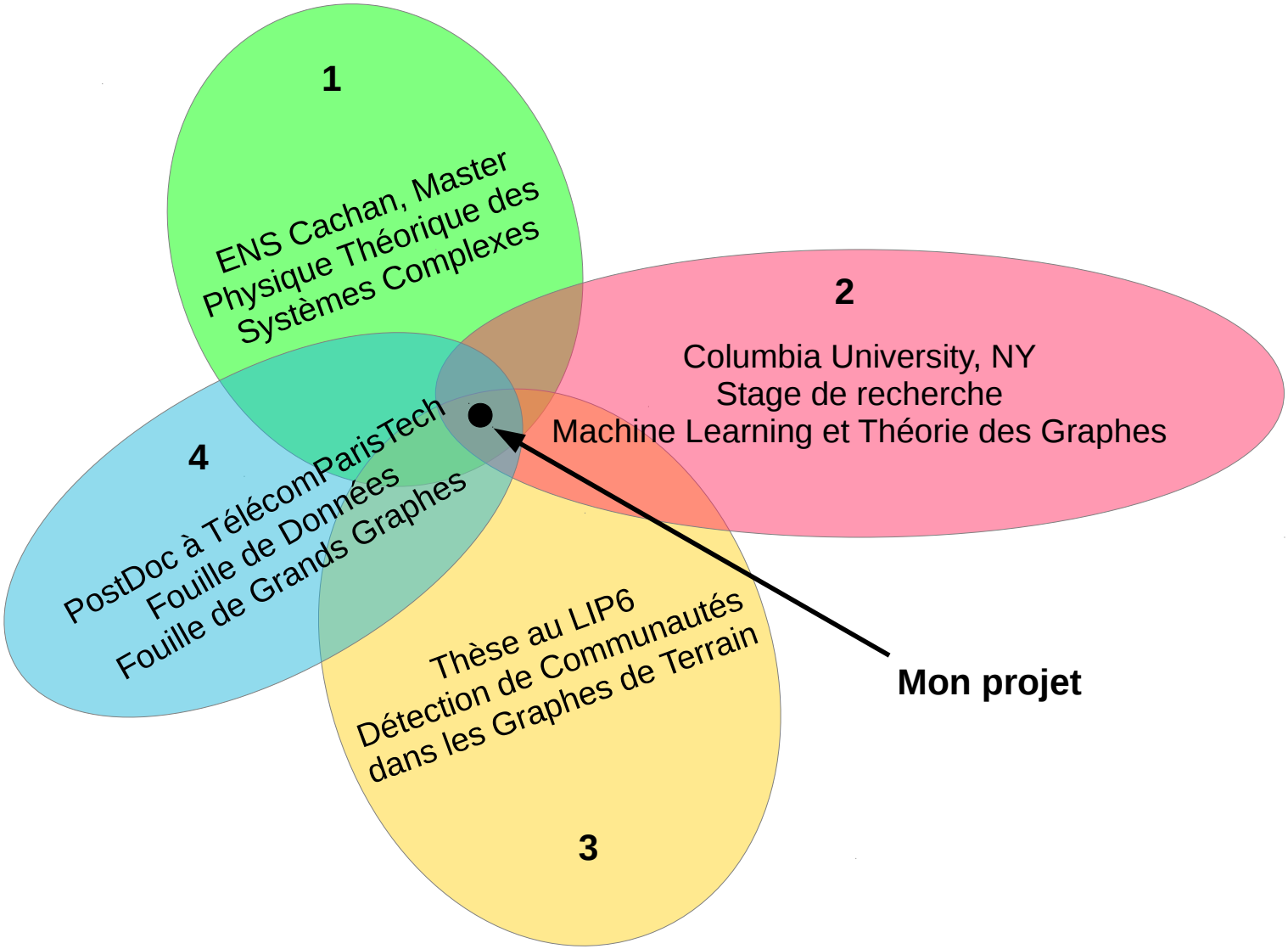
- graphes artificiels avec vérité de terrain
- graphes réels avec vérité de terrain
- Wikipédia (pages et liens hypertextes + catégories)

Conception d'algorithmes pour

- trouver toutes les communautés d'un nœud
- trouver toutes les communautés d'un graphe
- compléter un ensemble de nœuds en une communauté

▶ **Approche compétitive et originale, orthogonale aux optimisations de fonctions de qualité de l'état de l'art**

Mon parcours



Production scientifique

- **4 journaux, 8 conférences, 9 autres publications**
- Plus de 20 coauteurs, riches collaborations internationales

Sélection de publications

- Large Scale Density-friendly Graph Decomposition via Convex Programming. Danisch, Chan et Sozio. **WWW2017**. A*.
- Local Triangle-densest Subgraphs. Samusevich, Danisch et Sozio. **ASONAM2016**. 13.6%.
- Learning a Proximity Measure to Complete a Community. Danisch, Guillaume et Le Grand. **DSAA2014**. 11%.

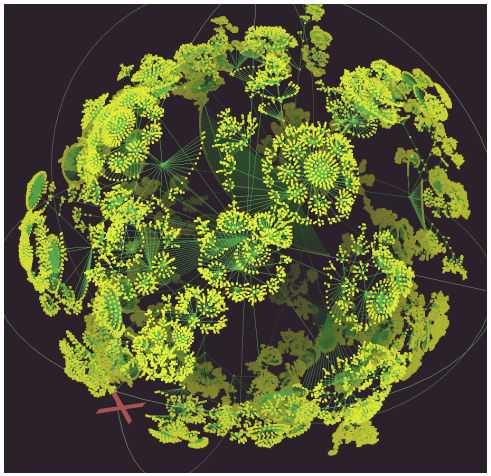
Développement logiciel important

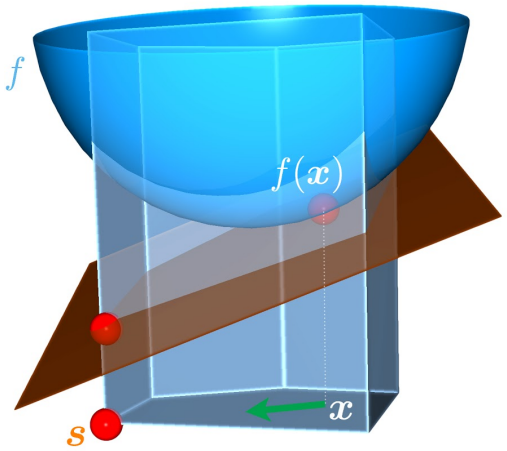
- **Code en C :** `https://github.com/maxdan94`
- **Twitter APP :** `http://bitly.com/socialK`
- **Twitter APP :** `http://bit.ly/easymention`

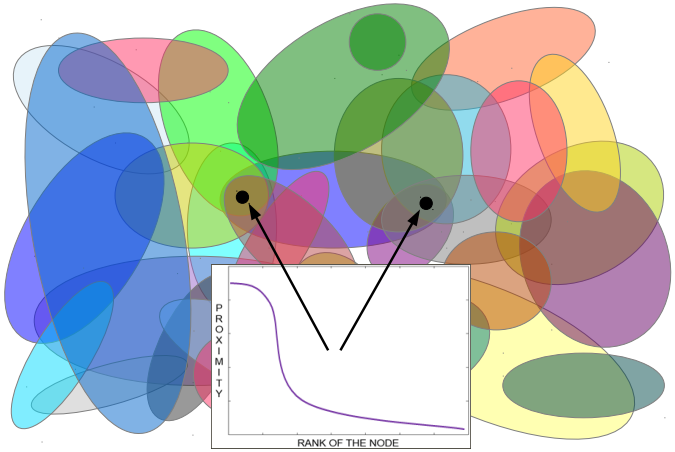
Intégration dans trois équipes CNRS

- ComplexNetworks - **LIP6**
  - Étude des graphes de terrain
  - Travaille avec des jeux de données réels et élabore des algorithmes efficaces en pratique
  - Énumération de triangles
- Optimisation combinatoire, algorithmique - **LAMSADE**
  - Forte expertise en optimisation combinatoire
  - Forte expertise en complexité paramétrée
  - Projet “Algorithmique pour les Masses de Données”
- Algorithmique distribuée et graphes - **IRIF**
  - Forte expertise théorique
  - Forte expertise en décompositions de graphes
  - Algorithmes conçus pour des familles de graphes

- **Projet :** Une algorithmique des graphes de terrain
- **Parcours :** ENS Cachan, Stage 1 an à Columbia University, Thèse au LIP6, PostDoc à TélécomParisTech.
- **Production scientifique :** 4 journaux, 8 conférences, 9 autres publications, développement logiciel important
- **Collaborations :** plus de 20 coauteurs, riches collaborations internationales
- **Intégration :** LIP6, LAMSADE et IRIF







SLIDES DE BACKUP

Algorithm 1 Parallel algorithm for finding $3, 4, 5$-cliques

$\eta \leftarrow$ Core ordering of G ▷ $O(m)$
for each node u in G **do**
  $\Delta_\eta(u) \leftarrow$ sorted list of neighbors v of u, $\eta(v) > \eta(u)$ ▷ $O(m)$
for each edge $(u, v) \in E(G)$ **do** ▷ parallel loop $O(m)$
  $\Delta_\eta(u, v) \leftarrow \Delta_\eta(u) \cap \Delta_\eta(v)$ ▷ $O(c)$
  for each w in $\Delta_\eta(u, v)$ **do** ▷ $O(N_3)$
    output triangle $\{u, v, w\}$
    $\Delta_\eta(u, v, w) \leftarrow \Delta_\eta(u, v) \cap \Delta_\eta(w)$ ▷ $O(c)$
    for each x in $\Delta_\eta(u, v, w)$ **do** ▷ $O(N_4)$
      output 4-clique $\{u, v, w, x\}$.
      $\Delta_\eta(u, v, w, x) \leftarrow \Delta_\eta(u, v, w) \cap \Delta_\eta(x)$ ▷ $O(c)$
      for each y in $\Delta_\eta(u, v, w, x)$ **do**
        output 5-clique $\{u, v, w, x, y\}$

Algorithm 1 requires $O(c \cdot \sum_{l=2}^{k-1} N_l)$ total number of operations.

k-cliques: datasets

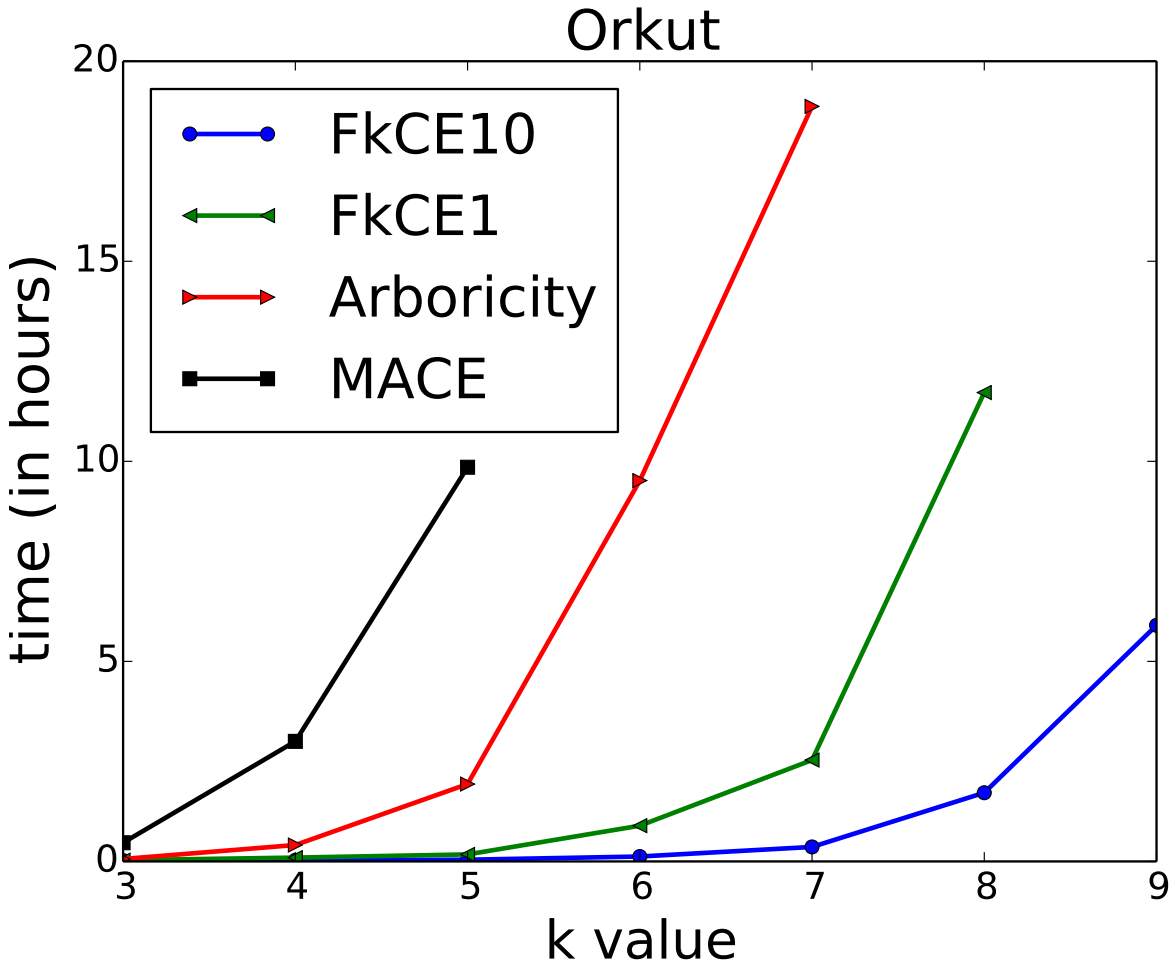
Table 1 : Our set of large graphs (we list all cliques)

networks	n	m	c	k_{max}	$N_{k_{max}}$
soc-pocket	1,632,803	22,301,964	47	29	6
loc-gowalla	196,591	950,327	51	29	2
Youtube	1,134,890	2,987,624	51	17	2
cit-patents	3,774,768	16,518,947	64	11	2
zhishi-baidu	2,140,198	17,014,946	78	31	4
WikiTalk	2,394,385	4,659,565	131	26	141

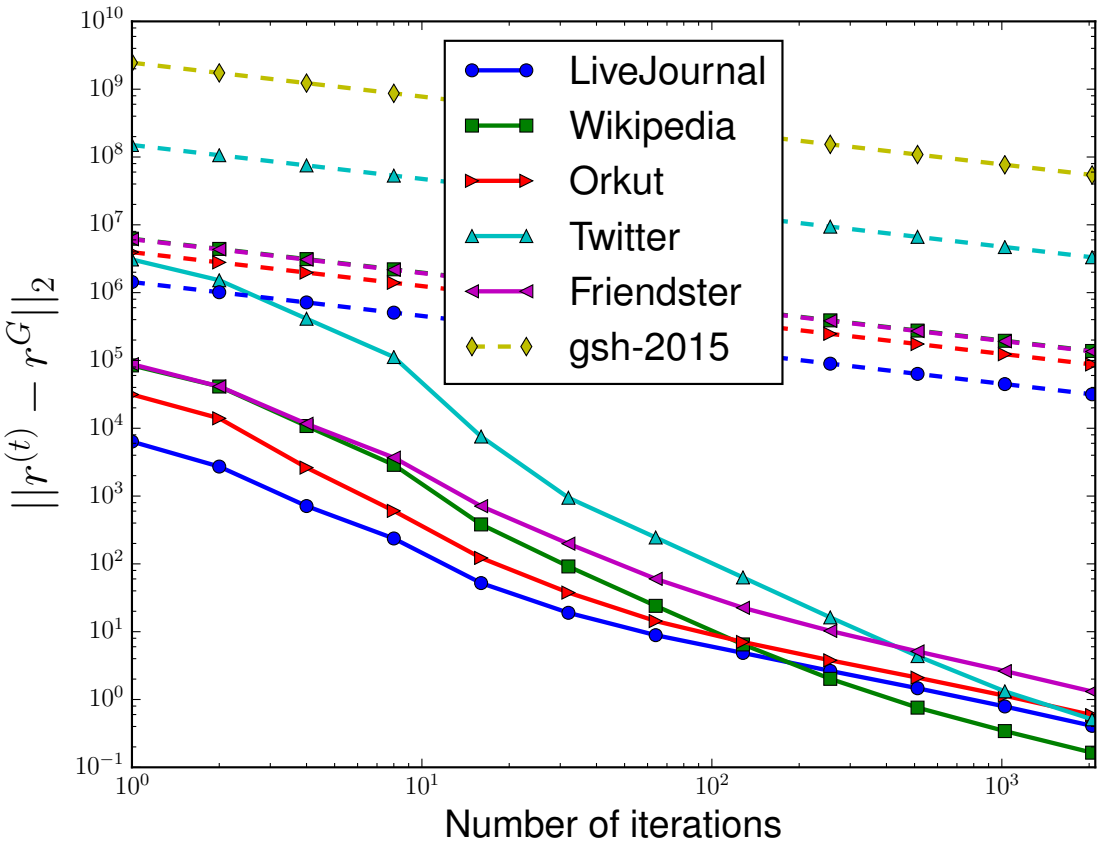
Table 2 : Our set of very large graphs (we list all k-cliques of limited size).

networks	n	m	c
DBLP	425,957	1,049,866	113
Wikipedia	2,080,370	42,336,692	208
Orkut	3,072,627	117,185,083	253
Friendster	124,836,180	1,806,067,135	304
LiveJournal	4,036,538	34,681,189	360
Twitter	52,579,683	1,614,106,500	2647

k-cliques: comparison to other methods, in practice

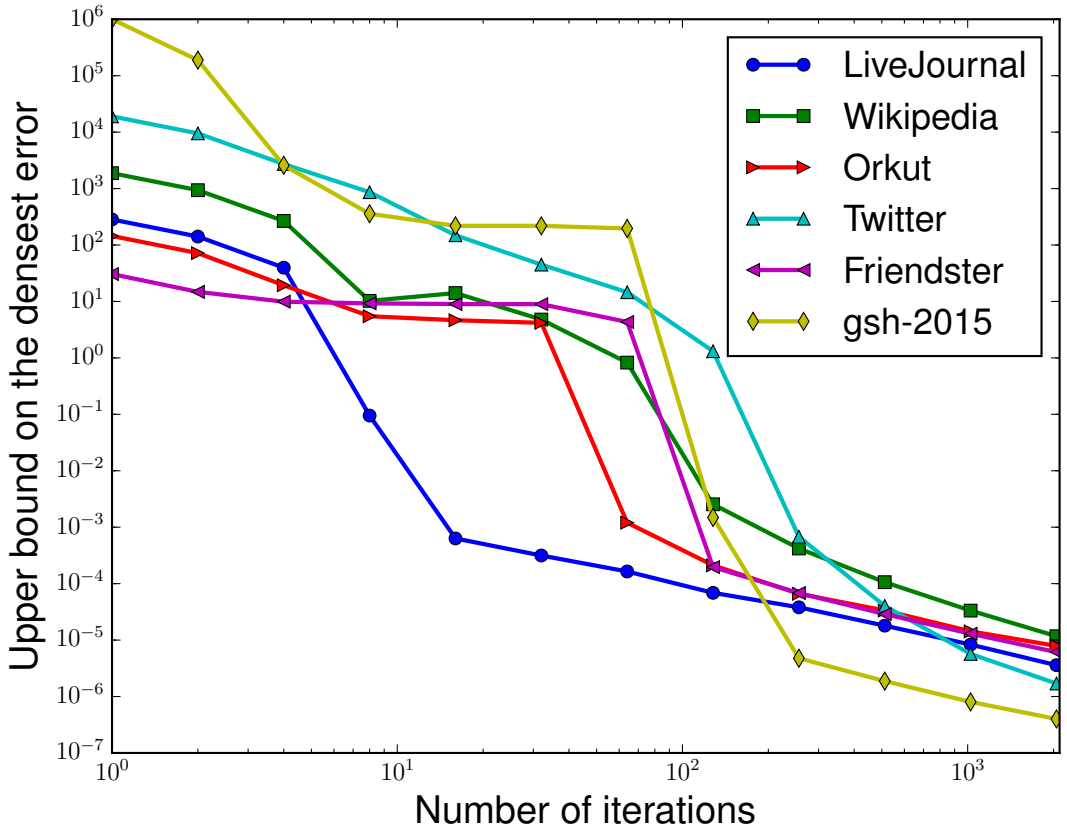


Density-friendly: Convergence of the **r** vector



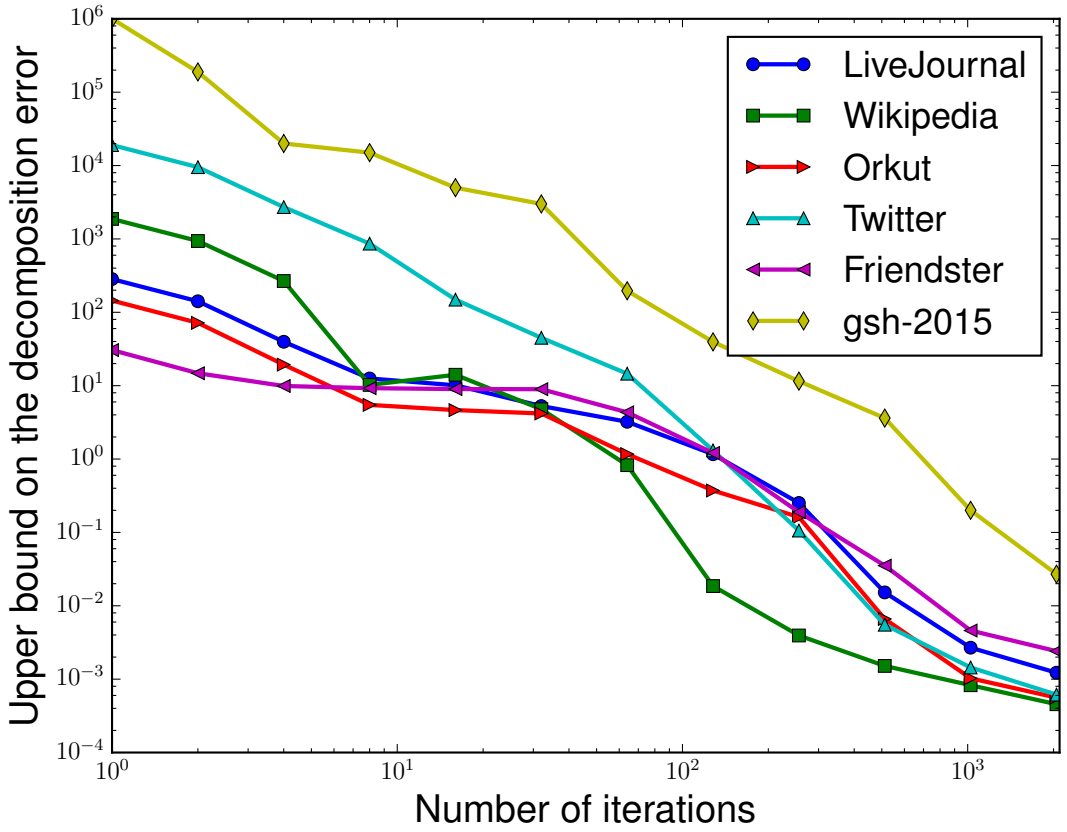
The convergence is in practice much faster than the worst case one.

Density-friendly: Densest multiplicative error



We obtain a 10^{-3} approximation of the densest subgraph within 300 iterations on all networks.

Density-friendly: Decomposition multiplicative error



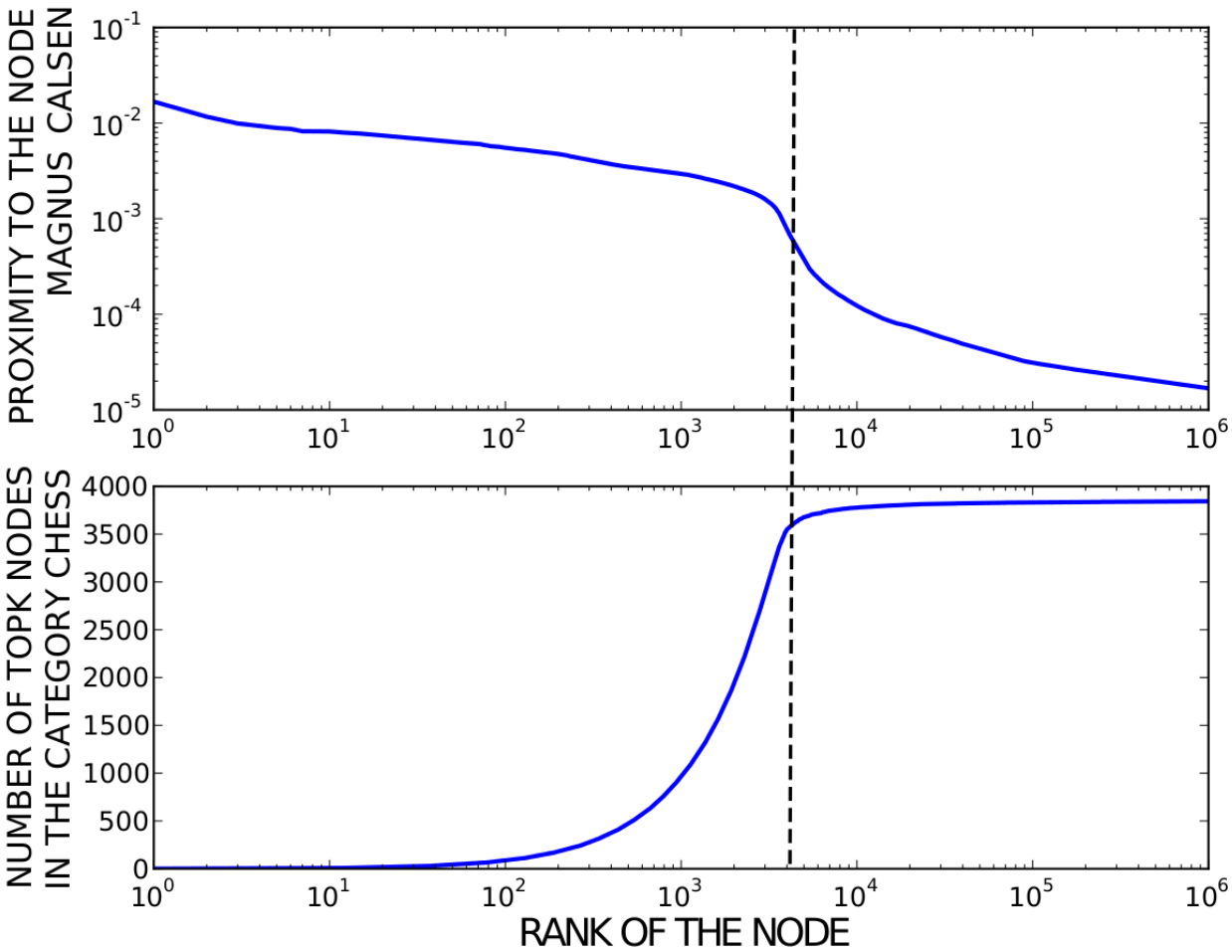
We obtain a 10^{-2} approximation of the full decomposition within 1000 iterations on all networks except gsh-2015 (almost 10^{-1}).

Collaborations

- Collaborations internationales :
  - Huber Chan à Hong Kong university. Spécialiste de l'algorithmique des graphes et de l'optimisation convexe.
  - Bivas Mitra à l'IIT Kharagpur. Spécialiste de l'étude des réseaux complexes.
  - Hernan A. Makse au City University of New York. Spécialiste de l'étude des réseaux complexes.
    Startup : `http://www.kcore-analytics.com/`.
  - Tony Jebarra à Columbia University à New York. Spécialiste en Machine Learning et graphes parfaits.
    Directeur de la recherche en Machine Learning chez NetFlix.
- Multiples collaborations nationales : plus de 20 coauteurs (L3i La Rochelle, LIFO Orleans, LIUM Le Mans, CRI Paris 1,CÉDRIC CNAM Paris, le2i Dijon).

Thèse. Communautés locales et mesures de proximité

Validation dans Wikipedia



Thèse. Communautés locales et mesures de proximité

Validation dans Wikipedia