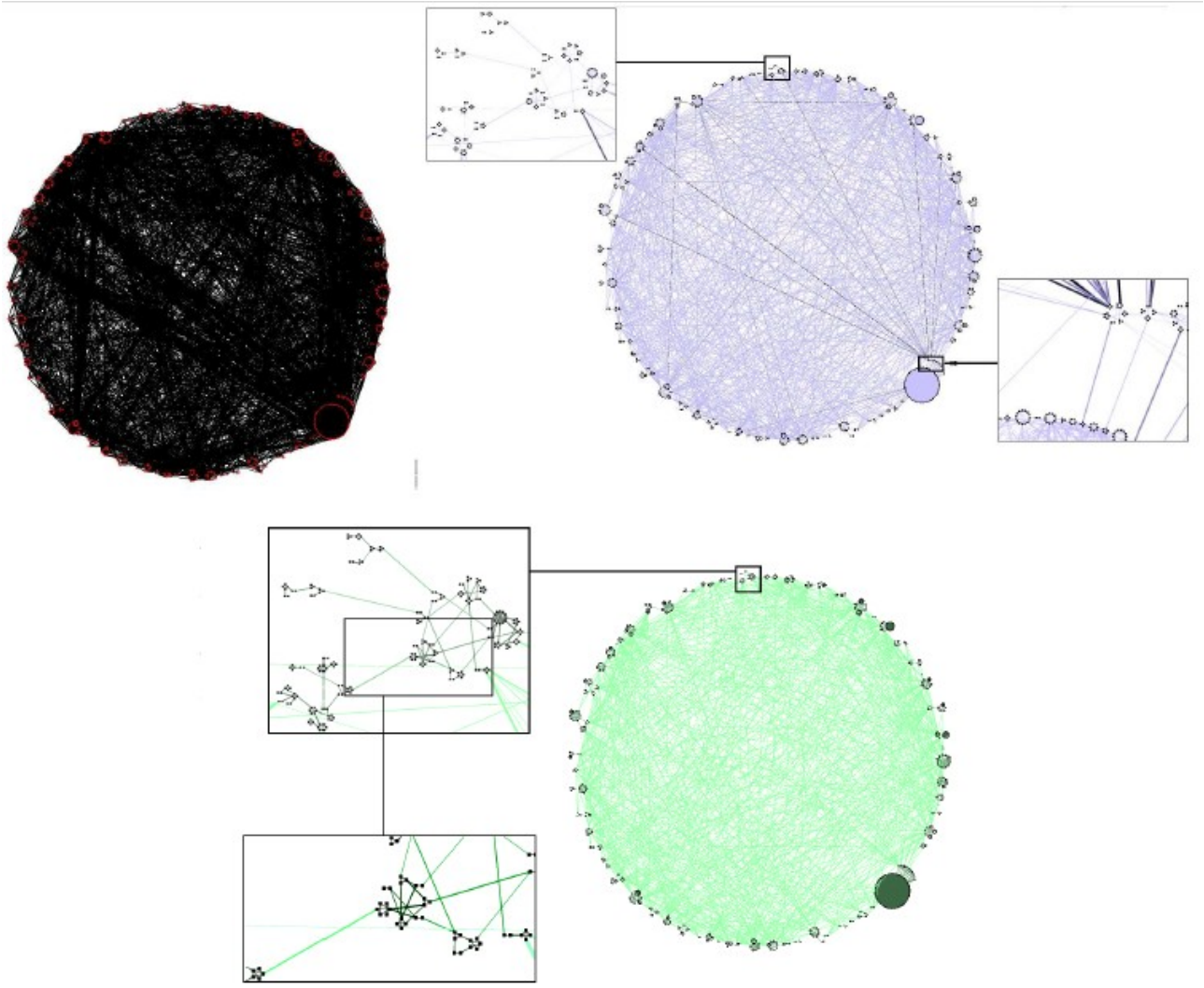
Graph clustering

Mathieu Barthelemy ENS de Lyon



Résumé

La découpe en communautés (graph clustering) joue un rôle clef, non seulement en informatique, mais aussi par ses applications dans de nombreux domaines scientifiques: telles les voies métaboliques en biologie ou la découpe en communautés en sociologie. Cette découpe peut être faite en rassemblant chaque nœud ayant un lien avec au moins la moitié des autres dans une même communauté[1]. La découpe ainsi obtenue donne un diamètre de deux dans chaque communauté. Cette découpe permet de plus une optimisation basée sur une bonne gestion des sommets isolés, ce qui permet de réduire le temps de calcul.

Introduction

La découpe en communautés (graph clustering) consiste à découper et à identifier, au sein d'un graphe, les ensembles de nœuds fortement connectés pour les regrouper au sein d'une communauté.

La découpe en communauté doit permettre de regrouper les nœuds fortement connectés entre eux et de séparer les nœuds peu connectés entre eux.

Ici, une méthode utilise la notion de haute connectivité d'un nœud. L'algorithme, ses propriétés et ses optimisations sont présentés dans le contexte des découpes en communautés.

Définitions

Dans ce qui suit G(V,E) est un graphe avec V sommets et E arrêtes.

Communautés fortement connectées

Une communauté fortement connectée est un sous graphe G'(V',E') de G dans lequel la coupe minimale est constituée d'au moins V'/2 arêtes.

Une autre définition, qui lui sera ici comparée, est basée sur le degré minimal des sommets de la communauté. Une communauté sera alors un sous graphe G'(V',E') de G dans lequel chaque sommet est relié à au moins V'/2 sommets.

Coupe

Une coupe constitue une partition de G(V,E) en deux ensembles V' et V''=V\V'. Les arêtes reliant ces 2 ensembles sont les arêtes à retirer pour effectuer une coupe. La coupe correspond donc à l'ensemble des arêtes ayant une extrémité dans V' et l'autre dans V''.

Le diamètre d'un graphe

Soit C l'ensemble des distances entre deux sommets quelconques du graphe G. Le diamètre de G est le plus grand élément de C.

Coupe minimale

Une coupe minimale est une coupe, qui dans un graphe non pondéré, minimise le nombre d'arêtes retirées pour séparer le graphe en deux.

Algorithmes

L'algorithme des sous graphes hautement connectés

L'algorithme

L'algorithme consiste à diviser le graphe tant que les sous graphes ne sont pas hautement connectés.
Une coupe minimale est alors utilisée. La coupe la plus utilisée est celle de Gomory-Hu [4].

Entrée

Un graphe G(V,E)

Sortie

Un graphe G(V,E) avec une découpe en communautés

Algorithme de découpe en communautés

Tant que le graphe est non vide:

- retirer tous les graphes fortement connectés
- si le graphe est non vide faire une coupe minimum dans chaque composante connexe

Retourner les communautés obtenues

Complexité

Soit N le nombre de communautés. La boucle de l'algorithme va être itérée au plus 2N-1 fois, ce qui correspond au nombre de nœuds d'un arbre binaire.
Au niveau des feuilles (communautés trouvées) la coupe ne sera pas réalisée. Ce qui correspond à N éléments. La coupe est donc réalisée pour N-1 éléments.
Le coût d'une coupe minimale est en O(v*e) [4] où v est le nombre de sommets et e est le nombre d'arêtes. Le coût total est donc borné par N + N-1(v*e). Pour la coupe minimale il existe un algorithme probabiliste en $O(e*\log^3 v)$ [1;4].
Ce qui fait alors une complexité en $N + N-1(e*\log^3 v)$.

Exemple

Le graphe suivant donne un exemple de l'utilisation de l'algorithme sur un graphe de 12 sommets. G n'est pas une communauté car il existe une coupe (en pointillé) utilisant seulement 2 arêtes ce qui est inférieur à V/2 (12/2). Il en est de même pour G2. G1, G3 et G4 sont des communautés. L'algorithme de coupe est ici utilisé 2 fois: sur G et G2.

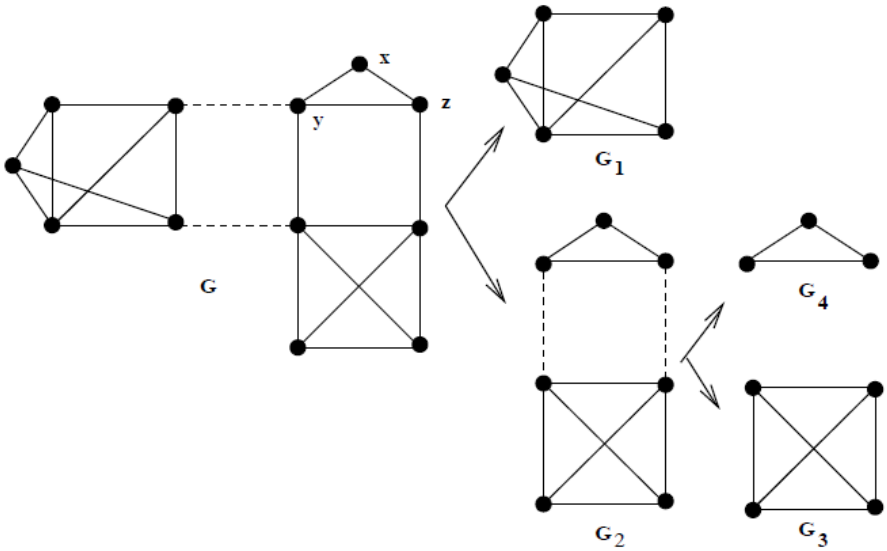


Figure 1: G est le graphe de départ. La découpe obtenue en communautés est G1; G3; G4.

Propriétés des graphes obtenus

Le diamètre d'un graphe est au plus 2.

Cette propriété peut être démontrée par l'absurde.
Supposons qu'il existe au sein d'une composante fortement connectée 2 nœuds à une distance de 3.
Alors, le premier nœud doit être lié uniquement à des nœuds qui ne sont pas connectés au second. Par définition d'une composante fortement connectée et comme le degré minimal d'un graphe est inférieur ou égal à la coupe minimale, chaque sommet est de degré au moins V/2. Le premier sommet est lié à au moins n/2 nœuds. Il reste alors n-2 nœuds non connectés au premier. Or le second doit se connecter lui aussi à n/2 nœuds non connectés au premier. C'est impossible car il manque au moins 2 nœuds. Le diamètre d'un graphe est donc au plus de 2.

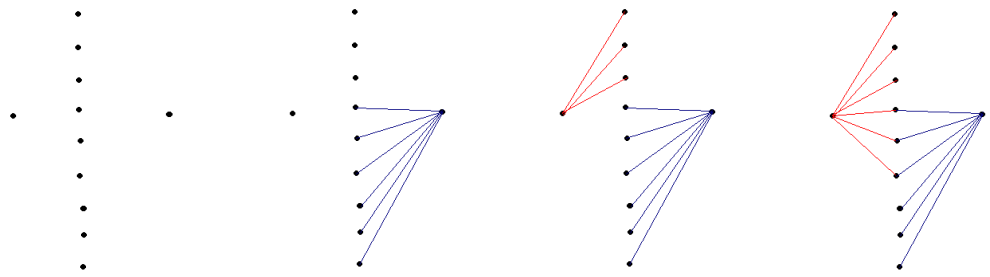


Figure 2: Démonstration graphique: le diamètre d'une composante fortement connectée est au plus 2. Deux nœuds ont au moins 2 sommets en commun.

Le nombre d'arêtes d'un graphe hautement connecté est quadratique.

Dans une composante fortement connectée, chaque nœud est connecté à au moins la moitié des autres nœuds. Chaque arête étant ici comptée 2 fois, cela donne V*V/2*1/2=

$V^2/4$. Le nombre d'arêtes d'un graphe hautement connecté est donc quadratique.

Une coupe minimale retire un nombre linéaire de sommets.

Une coupe sépare un graphe en deux parties. Une des deux parties peut être constituée d'un seul sommet. La coupe minimale est la coupe qui retire le moins de sommets pour couper le graphe en deux.
Dans un graphe, il existe dans le pire des cas, un sommet qui est relié à tous les autres. Une coupe retirant ce seul sommet retire V-1 arêtes. Toute coupe retirant plus de sommets n'est donc pas minimale.

Améliorations possibles

Tester les différentes découpes possibles

Le choix glouton pour le minimum cut ne conduisant pas toujours à une découpe en communautés optimale, il peut être intéressant de répéter plusieurs fois l'algorithme en effectuant des choix non localement optimaux.

Retirer les sommets de très faible degré

Le temps de calcul peut être nettement amélioré en retirant tous les sommets ayant un très faible degré. Cela permet d'éviter de nombreuses itérations très couteuses de l'algorithme.

Rattacher les sommets isolés aux communautés existantes

Les éléments isolés peuvent être rattachés à la communauté avec laquelle ils ont le plus de similarités.

Cette idée peut être répétée plusieurs fois pour obtenir ce nouvel algorithme.

L'algorithme avec les améliorations

Entrée

Un graphe G(V,E)

Sortie

Un graphe G(V,E) avec une découpe en communautés

Algorithme de découpe en communauté gérant les sommets isolés

Tant que G est non vide
- retirer tous les graphes fortement connectés
- retirer tous les sommets avec un faible degré
- si le graphe est non vide faire une coupe minimum dans chaque composante connexe
- sinon rattacher les sommets isolés aux différentes communautés
- retourner les communautés obtenues

Complexité

Soit I le nombre de sommets isolés et R le coût pour détecter, retirer et réinsérer un

sommet isolé.

Le coût total est donc borné par N-I + N-I-1(v*e) +IR. Pour la coupe minimale probabiliste cela fait alors une complexité en N-I + N-I-1(e*$\log^3$v)+ IR.

Avec la deuxième définition de fortement connecté

Les graphes obtenus seront très similaires en utilisant la définition de fortement connecté avec le degré minimal égal à V/2. Ils ont les mêmes propriétés que ceux obtenus avec la définition utilisant une coupe minimale égale à V/2 car la condition être de degré minimal supérieur à V/2 implique d'avoir une coupe minimale de V/2. Une implémentation permettrait de comparer les découpes obtenues avec les 2 définitions de fortement connecté. Il est peut être plus facile de détecter un sommet de degré inférieur à V/2 qu'une coupe de V/2.

Evaluation de la découpe en communautés

La qualité des découpes obtenues par les algorithmes de découpes en communautés est généralement évaluée par des fonctions dites de qualité. Il en existe de nombreuses [3a].

La modularité

La modularité Qm est le rapport entre les arêtes internes à la communauté divisé par: la moitié du degré moyen du graphe multiplié par le nombre de sommets du sous graphe.

La performance

La performance mesure le taux d'arêtes bien placées selon la découpe en communautés. Cela correspond au nombre d'arêtes entre 2 sommets d'une même communauté, plus le nombre d'arêtes absentes entre les sommets de communautés différentes, le tout divisé par le nombre d'arêtes d'une clique ayant le même nombre de sommets que le graphe.

Visualisation des graphes de communautés

La découpe en communautés est une étape essentielle pour recueillir l'information des grands graphes de terrain. La deuxième étape correspond à la visualisation. 3 algorithmes de visualisation ont été notamment développés.

SWIF

SWIF est un logiciel développé au LABRI de Bordeaux. Les communautés sont représentées sous formes de cercles concentriques [7].

Représentation des communautés en cercles

L'idée principale est de représenter les communautés sur un cercle [7]. Une sous communauté y est alors représentée de manière récurrente sous forme d'un nouveau cercle qui prend son centre sur le cercle de niveau supérieur.

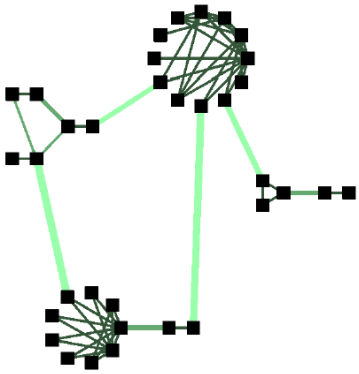


Figure 4: Un exemple de l'utilisation de l'algorithme de visualisation avec des cercles.

Représentation thermique

Cette représentation utilise un modèle de physique [5].

Les nœuds d'une communauté ont des arêtes qui représentent une force attractive à longue distance. Mais pour éviter le chevauchement des nœuds sur le graphique, les nœuds ont aussi des forces répulsives. Le système se comporte alors comme un oscillateur qui va osciller de manière infinie dans une zone limite. Pour stabiliser le système: des forces de frottement simulent l'effet d'un abaissement de la température et donnent une répartition qui permet une bonne visualisation du graphe.

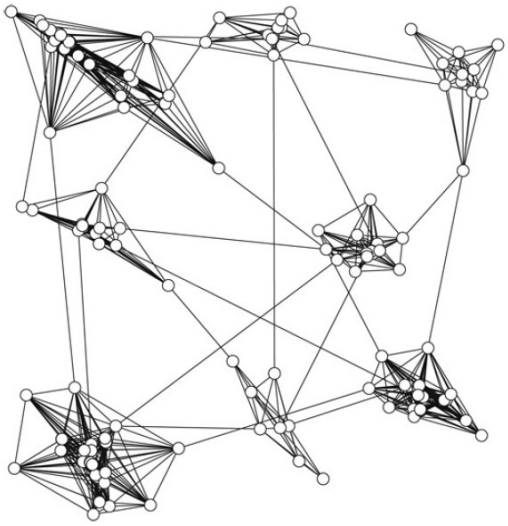


Figure 5: Un exemple de l'utilisation de l'algorithme de visualisation GEM [5].

Applications

La découpe en communautés trouve ses applications dans de nombreux domaines scientifiques. Les applications les plus connues se retrouvent dans les domaines sociaux [8] et en biologie [9].

Systèmes biologiques.

Les graphes sont très fréquemment utilisés en biologie; les exemples les plus connus concernent les réseaux métaboliques et les réseaux protéiques.

La protéomique

Le protéome correspond à l'ensemble des protéines d'une cellule. Les protéines sont des molécules interagissant les unes avec les autres. Les protéines dans la cellule sont regroupées sous forme de complexes fonctionnels. La découpe en communautés permet ici de découvrir ces complexes et d'en comprendre le rôle et les régulations. La compréhension de ces complexes a un rôle clef dans les maladies neurodégénératives et dans les processus cancéreux.

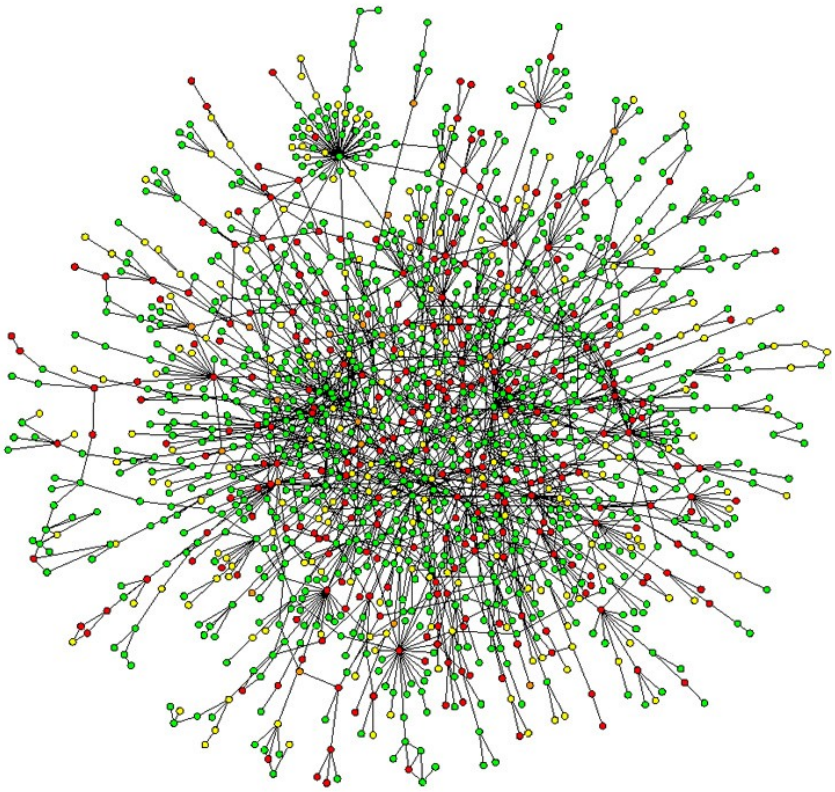


Figure 6: Le protéome humain montre qu'une découpe en communautés est nécessaire à une compréhension des graphes d'interactions protéine-protéine

Le métabolisme

Les voies métaboliques sont des ensembles de réactions chimiques catalysées par des enzymes. Les produits d'une réaction sont les substrats d'une autre réaction. Ces liens produits-substrats permettent de construire un graphe. La découpe en communautés permet de comprendre le rôle de certaines enzymes et de mieux comprendre les pathologies génétiques liées au dysfonctionnement de ces enzymes et de prévenir d'autres pathologies en comprenant mieux les effets de l'alimentation au sein de ce graphe.

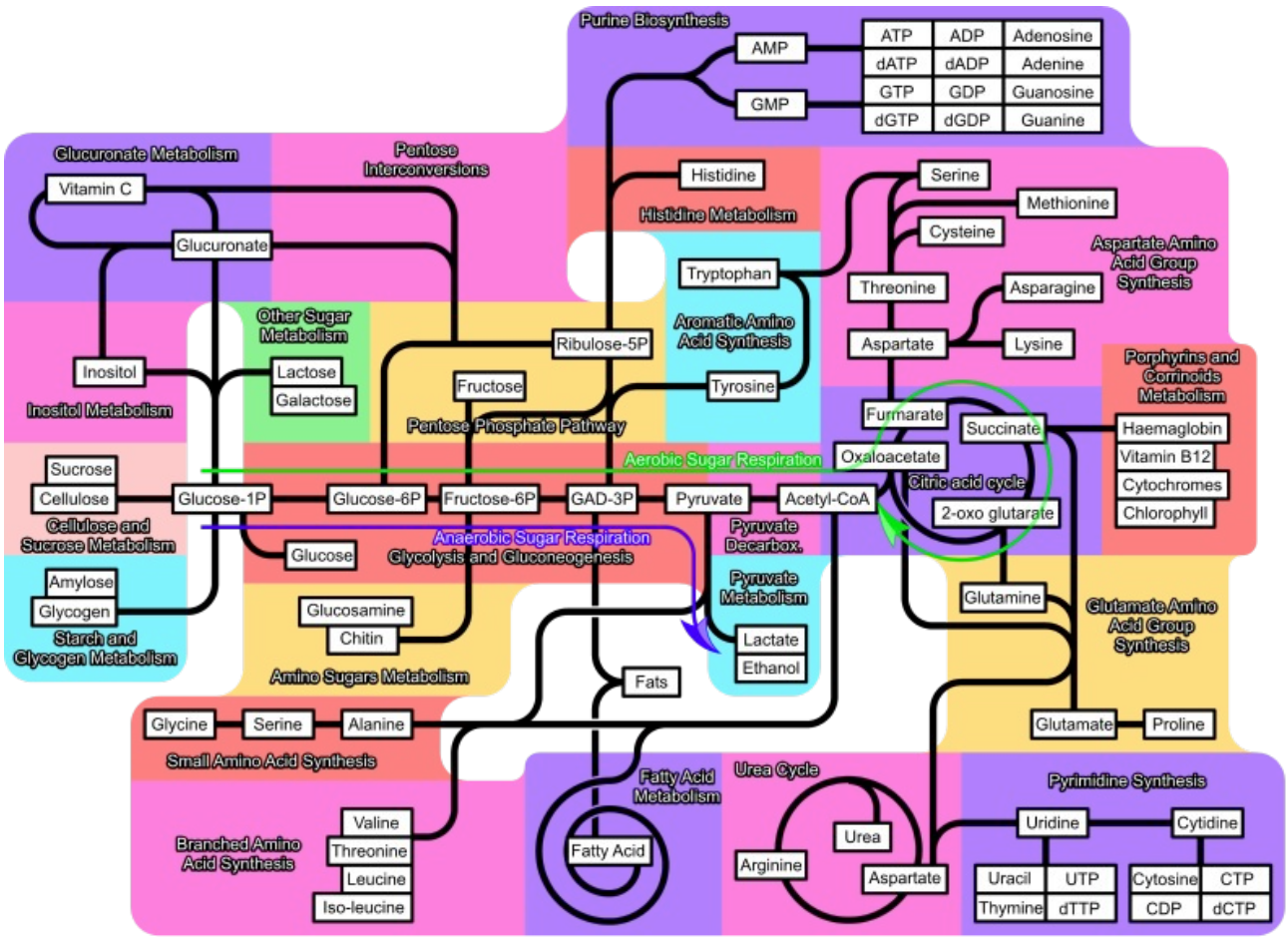


Figure 7: les principales voies métaboliques d'une cellule sont regroupés par les biologistes en communautés.

Modèles épidémiologiques

La transmission des maladies nosocomiales au sein des CHU est un enjeu majeur de santé publique [7]. De manière évidente: la transmission de ces maladies dépend de la structuration en sous communautés. Les algorithmes de clustering permettent donc de mieux évaluer et de mieux cibler la prévention au sein des services hospitaliers [9].

Réseaux sociaux

Le réseau velov de Lyon

Le réseau velov désigne un système de Vélos en libre service mis en place dans la communauté urbaine de Lyon depuis mai 2005 [8]. Les bornes où les vélos sont pris constituent les sommets d'un graphe. Les trajets entre 2 bornes constituent une arête. Comprendre quelles sont les communautés de sommets permet de mieux gérer ce réseau en évitant des pénuries ou des saturations des bornes. C'est un projet de l'IXXI (Institut des Systèmes Complexes Rhone-Alpin).

Réseau de téléphonie mobile

Un graphe des appels en Belgique a été réalisé [7]. La découpe en communautés permet de comprendre, de visualiser les groupes sociaux et de leur proposer des forfaits adaptés à leurs besoins.

Conclusion

La découpe en communautés est utilisée dans de nombreux domaines scientifiques et permet de distinguer l'information du bruit. Les algorithmes existants sont performants et permettent de manipuler des graphes particulièrement grands tels que le réseau ARXIV [7] et les réseaux biologiques. Des algorithmes de visualisation permettent alors d'explorer ces graphes.

Bibliographie

Bibliographie primaire

[1] A clustering algorithm based on graph connectivity. E Hartuv et Ron Shamir
C'est l'article de départ de ce travail.

Bibliographie secondaire

articles généraux

[2] http://www.elsevier.com/authored_subject_sections/P05/misc/Schaeffer.pdf
Cette revue a servi de base pour ce rapport. Elle est particulièrement complète. Avec plus de 200 références bibliographiques.

[3] http://www.grappa.univ-lille3.fr/~candillier/clusters/these.pdf
Cette thèse offre un aperçu plus général sur le clustering. Notamment les méthodes hors théorie des graphes y sont présentées.

[3a] Pascal Pons Détection de communautés dans les grands graphes de terrain 2007
http://psl.pons.free.fr/publi/these_pascal_pons.pdf
Une thèse qui traite globalement le sujet et qui fait un très bon état de l'art sur les fonctions de qualité.

[4] Algorithmes d'approximation, Vazirani, V. Collection IRIS Springer 2006
isbn:9782287006777
Ce livre présente plusieurs algorithmes de coupe minimale avec différentes approximations.

[5] A fast adaptive layout algorithm for undirected graphs A Frick, A Ludwig, H Mehldau - Graph Drawing, 1995
Algorithme d'affichage de la structure d'un graphe en communautés utilisant un modèle mimant la cristallisation en physique.

Travaux au sein de l'ENS sur ce thème

[6] http://www.ens-lyon.fr/DSM/SDMsite/M2/stages_M2/Tabourier.pdf
Un rapport de stage de M2 d'un étudiant de physique ayant travaillé sur les réseaux sociaux dont internet.

[7] Visualisation des communautés au sein de grands graphes de terrain. Lucie Martinet
Un rapport de L3 sur la visualisation des communautés.

[8] Modélisation statistique cyclique des locations de Vélo’v à Lyon
Pierre BORGNAT, Patrice ABRY, Patrick FLANDRIN archives ouvertes.

[9] Le projet mosar https://www.mosar-sic.org/mosar/en-GB/
Le projet mosar est un projet de modélisation de la transmission des maladies nosocomiales au sein des CHU auquel participe l'IXXI de l'ENS de Lyon.