Université de Namur

Faculté d'informatique

Rapport académique : analyse de graphes

SDASM101 – Graph Mining

Graphe 1 : Primary school temporal network data

Graphe 2 : Hospital ward dynamic contact network

Abel Idrice ADJIEUFACK Certificat Data Science

Céline ARGYROPOULOS Certificat Data Science

David RUEN Certificat Data Science

18 décembre 2025

1 Présentation des articles SocioPatterns

1.1 Primary school - cumulative networks

Le réseau étudié est un graphe construit à partir d'un dataset collecté via SocioPatterns, représentant les interactions entre élèves et enseignants d'une école. Ce dataset contient le réseau temporel des contacts entre les enfants et les enseignants utilisé dans l'étude publiée dans BMC Infectious Diseases 2014. Les arêtes sont pondérées par le temps total de contact, de sorte que des liens plus épais indiquent des interactions plus longues. Les noeuds sont colorés par catégories [classes et teachers] et permettant de visualiser la structure communautaire de l'école. Cette représentation met en évidence à la fois les interactions au sein des classes et les contacts inter-classes, ainsi que le rôle central des enseignants [magenta] dans le réseau. D'après la Figure 1, 242 noeuds (nombre d'élèves par classe et nombre d'enseignants) et 8317 arêtes sont nécessaires pour décrire cette interaction au sein de cette école.

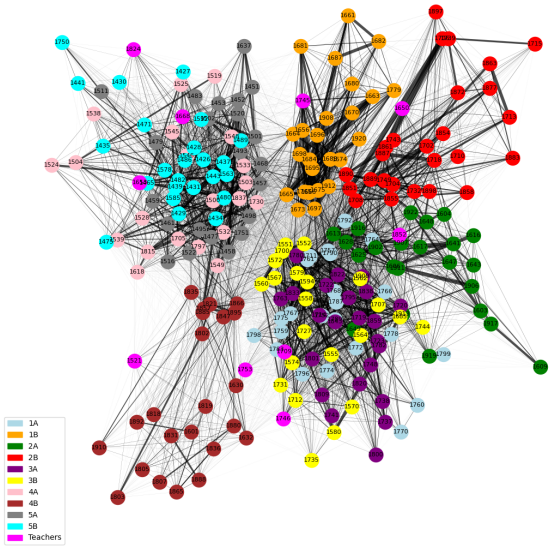

FIGURE 1 – Structure du graphe étudié.

1.2 Hospital ward dynamic contact network

Les données analysées proviennent de la base de données SocioPatterns et ont été collectées dans une unité de gériatrie d'un hôpital universitaire en France. Elles sont associées à l'article *Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors*. Les interactions entre individus ont été mesurées à l'aide de badges RFID portés sur la poitrine, permettant de détecter des interactions en face-à-face à une distance d'environ 1 à 1,5 mètre, avec une résolution temporelle de 20 secondes. Un "contact" entre deux individus est défini comme la détection d'au moins un échange de signal RFID au cours d'un intervalle de 20 secondes. Un contact est considéré comme continu tant que des échanges sont détectés dans des intervalles consécutifs. Le jeu de données contient des événements de contact horodatés entre paires d'individus (non identifiés), ainsi que leur rôle au sein de l'hôpital.

2 Article 1 : Primary school - cumulative networks

2.1 Propriétés structurelles

2.2 Distribution de degré

Le degré d'un sommet correspond au nombre de voisins. Ici, l'élève 1551 (classe 3B) présente le degré maximal (134). La densité est relativement faible (≈ 0.2852) : les interactions sont loin d'être complètes, mais restent suffisantes pour permettre une diffusion rapide. Enfin, la somme des degrés vaut $16\,634 = 2M$, cohérente avec un graphe non orienté et $M = 8\,317$ arêtes.

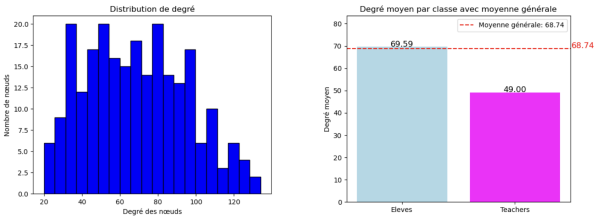

FIGURE 2 – Degré de distribution (a) et moyen par classe (élèves et teachers)

Elle indique que chaque individu n'interagit qu'avec une fraction limitée de la population. Cependant, la connectivité globale restant élevée, une propagation peut néanmoins être rapide.

2.3 Clustering

Le coefficient de clustering mesure la tendance des voisins d'un noeud à être connectés. Il vaut 0.5255 (Figure 3), indiquant une cohésion locale marquée, particulièrement dans certaines classes (p.ex. 2B ou 5B), ce qui peut accélérer des transmissions au sein de ces groupes.

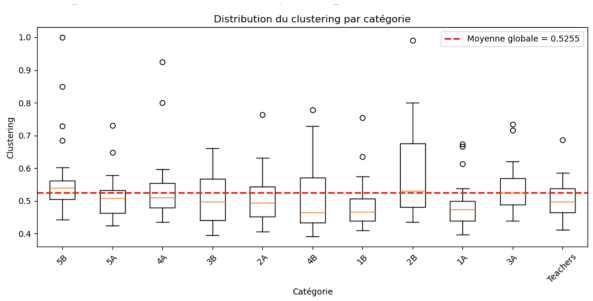

FIGURE 3 – Distribution du coefficient de clustering en fonction des catégories

2.4 Centralité

Les centralités (betweenness, closeness, Katz, PageRank) quantifient l'importance structurelle des noeuds. Le Tableau 1 montre que 1551 (3B) est maximal en betweenness, closeness et PageRank (médiation, accessibilité, influence globale), tandis que Katz met en avant 1743 (2B), davantage lié à l'influence via connexions indirectes.

	Betweenness	Closeness	Katz	PageRank
Valeur	0.01327	0.69253	0.14211	0.00740
Noeud	1551(3B)	1551(3B)	1743(2B)	1551(3B)

TABLE 1 – Valeurs des centralités et noeud le plus central selon chaque mesure.

D'après la figure 4, il y a de fortes corrélations statistiquement (rho supérieur à 0.9) significatives entre la closeness, la betweenness et le PageRank, indiquant qu'elles identifient globalement les mêmes noeuds centraux dans le réseau. En revanche, la centralité de Katz est modérément (Spearman) et négativement (Pearson) corrélée avec ces mesures, suggérant qu'elle capture une notion différente de centralité, davantage liée aux influences indirectes. Les coefficients de Spearman, légèrement supérieurs à ceux de Pearson, indiquent des relations monotones mais non strictement linéaires, tandis que les p-values très faibles confirment la robustesse de ces résultats.

FIGURE 4 – Mesures de centralité.

2.5 Comparaison avec des modèles de référence (Erdős-Rényi, Barabási-Albert, Watts-Strogatz)

Pour mieux comprendre la structure du graphe (Figure 1), il est utile de la comparer à celle de modèles aléatoires (Erdős-Rényi) et mécanistiques (Barabási-Albert et Watts-Strogatz), permettant d'évaluer si les propriétés (distribution des degrés, clustering et modularité) du réseau réel sont le fruit du hasard ou le résultat de mécanismes d'organisation spécifiques.

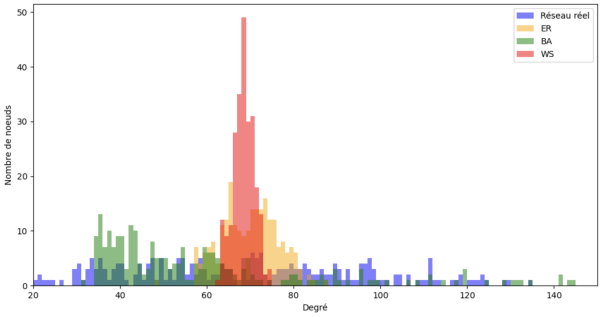

FIGURE 5 – Comparaison des distributions de degré.

Le réseau réel présente une modularité élevée (0,6684), cohérente avec une organisation par classes. Aucun des modèles ER, BA ou WS ne reproduit simultanément cette modularité et la distribution de degrés observée (Figure 5) : ER est trop homogène, BA accentue des hubs, et WS privilégie une structure locale.

2.6 Détection de communautés

Louvain détecte six communautés (Figure 6), alignées majoritairement sur les niveaux/classes. La modularité élevée (0.6684) confirme une organisation non aléatoire ; les enseignants apparaissent comme noeuds "relais" entre communautés, susceptibles de faciliter la transmission inter-classes.

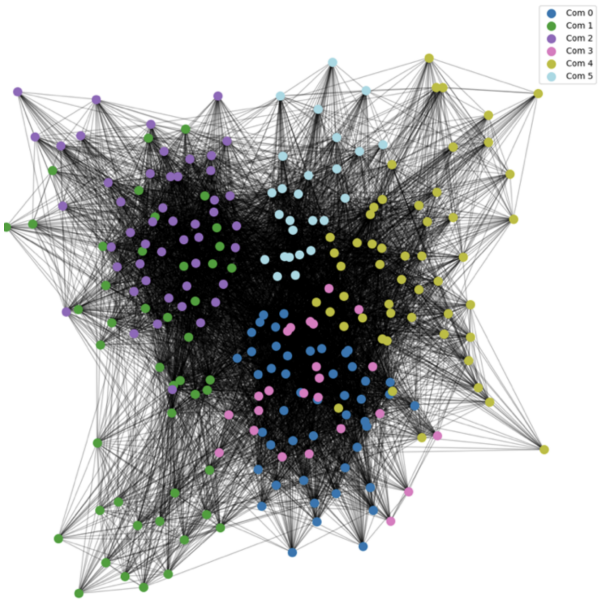

FIGURE 6 – Représentation des communautés d'après l'algorithme de Louvain

	Com 0	**Com 1**	**Com 2**	**Com 3**	**Com 4**	**Com 5**
Effectifs	48	46	24	47	50	26
Composition	23 (5B) 22 (5A) 2 (Teachers) 1 (4B)	1 (5B) 21 (4A) 22 (4B) 2 (Teachers)	23 (1A) 1 (Teachers)	22 (3B) 23 (3A) 2 (Teachers)	23 (2A) 26 (2B) 2 (Teachers)	25 (1B) 1 (Teachers)

TABLE 2 – Composition et effectifs des communautés détectées par l'algorithme de Louvain.

2.7 Matrice de contact entre classes et temporalité

Le jeu de données école est un réseau temporel échantillonné toutes les 20 secondes : chaque ligne correspond à un intervalle de proximité de 20 s. Sur l'ensemble de la période, on observe 125 773 événements de contact, soit 2 515 460 secondes (environ 698,7 heures) de proximité cumulée.

La Figure 7 présente la matrice des durées cumulées (en heures) entre catégories. La diagonale domine nettement, ce qui confirme que la majorité des interactions est intra-classe. Les contacts inter-classes les plus importants concernent surtout des classes parallèles d'un même niveau (p.ex. 3A-3B = 18,4, 2A-2B = 17,8). Les enseignants concentrent des liens inter-groupes, ce qui soutient leur rôle de ponts dans la diffusion inter-classes.

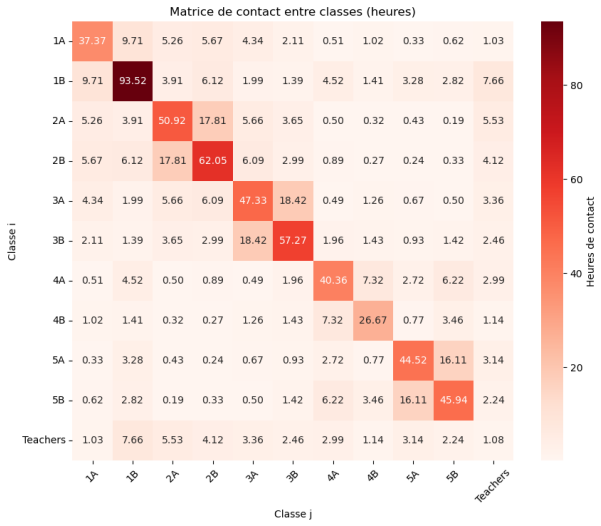

FIGURE 7 – Matrice des durées cumulées de contact entre catégories (heures).

2.8 Marche aléatoire

Une marche aléatoire consiste à se déplacer itérativement vers un voisin choisi au hasard, en privilégiant les voisins non visités. D'après la Figure 7, les deux marches explorent des zones différentes mais partagent 22 noeuds, suggérant des points de passage récurrents (noeuds bien connectés/charnières) importants pour la propagation. La longueur continue des trajectoires indique un réseau peu fragmenté : une contamination peut emprunter plusieurs corridors, avec des points de convergence à surveiller.

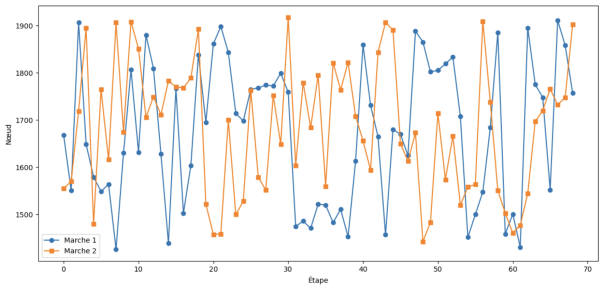

FIGURE 8 – Parcours de deux marcheurs aléatoires.

2.9 Processus épidémique

Pour étudier la propagation d'une épidémie dans le réseau scolaire, nous comparons la simulation SIR depuis le noeud central (1551) et un noeud périphérique (1609) afin de mesurer l'impact de la position structurelle et d'identifier les noeuds critiques (Figure 8).

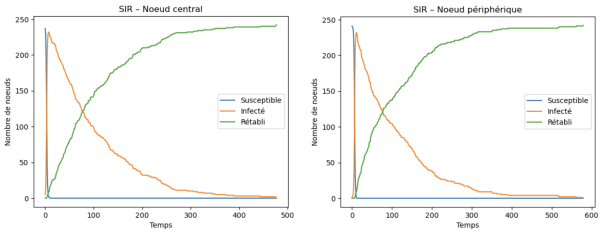

FIGURE 9 – Comparaison du pic d'infectés en fonction du noeud initial.

D'après cette figure, le pic d'infectés est très similaire quel que soit le noeud initial : 233 lorsqu'il démarre depuis le noeud central et 229 depuis le noeud périphérique. Cette faible différence suggère qu'en l'absence de mesures de mitigation, le réseau ne présente pas de barrières structurelles suffisantes pour contenir la propagation, même lorsque l'infection part d'un noeud marginal.

3 Article 2 : Hospital ward dynamic contact network

3.1 Représentation en graphe agrégé

Dans un premier temps, un graphe agrégé est construit à partir de ces données, où une arête entre deux individus indique qu'au moins un contact a eu lieu durant la période d'observation. Les arêtes peuvent être pondérées afin de prendre en compte l'intensité des interactions, par exemple via le nombre de contacts ou leur durée cumulée.

3.2 Propriétés statistiques

Dans cette section, le graphe agrégé non orienté est caractérisé à partir des données : taille, densité, connectivité, distribution de degré et clustering. Ces mesures fournissent une description globale de la structure du réseau avant d'étudier l'importance des noeuds (centralités) ou l'organisation en groupes (communautés).

Afin d'évaluer si certaines propriétés observées du réseau peuvent être expliquées uniquement par la distribution de degré, le graphe est comparé à un null model basé sur le configuration model.

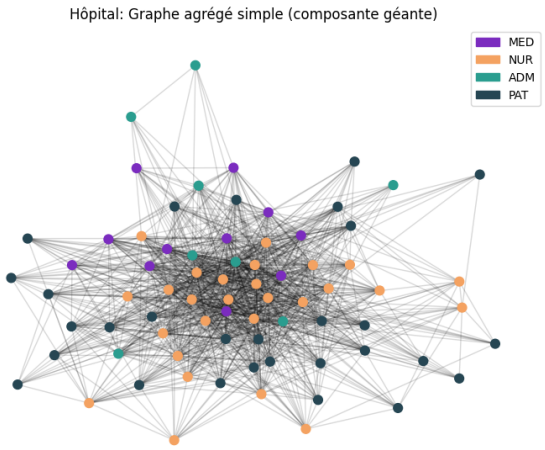

FIGURE 10 – Structure de graphe étudié (graphe agrégé simple).

Sur le graphe agrégé non orienté (arête si au moins un contact sur la période), on obtient N=75 noeuds et M=1 139 arêtes, soit une densité de 0,410. Le graphe est entièrement connexe (1 composante, composante géante = 100% des noeuds). La distribution de degré est centrée sur un degré moyen de 30.37 (médiane 27), avec un maximum de 61. Le clustering est élevé : clustering moyen (moyenne locale) = 0,640 et transitivity = 0,588.

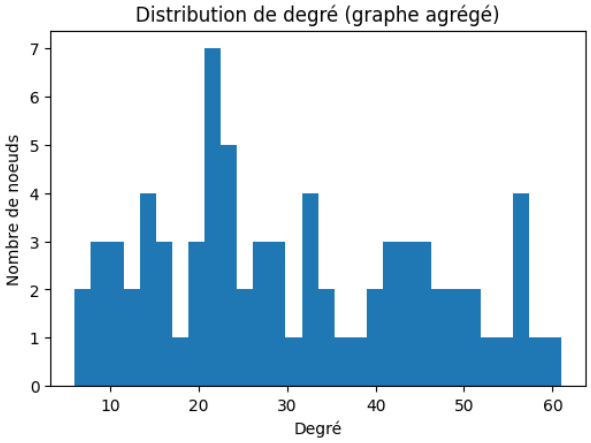

FIGURE 11 – Distribution de degrés.

Afin d'évaluer si le clustering observé peut s'expliquer uniquement par la distribution de degré, nous comparons le réseau à un null model basé sur le configuration model. Sur 30 réalisations, la transitivity moyenne du null model vaut environ 0,393 (+/- 0,007) et le clustering moyen local est d'environ 0,404 (+/- 0,009), nettement inférieurs aux valeurs observées (0,588 et 0,640). Cela suggère que la présence de triangles et de structure locale ne s'explique pas uniquement par la séquence de degrés.

3.3 Centralité

Dans l'article étudié, les "super spreaders" sont identifiés sur base du nombre et de la durée des contacts, ce qui correspond directement à une mesure de degré (ou de

strength dans le cas pondéré). La betweenness centrality capture une notion différente d'importance liée au rôle d'intermédiation dans le réseau, mais ne correspond pas à la définition dans l'article.

Le résultat de l'article est reproduit ici : l'existence de "super spreaders". En ne considérant que les contacts entre personnel soignant (HCW = NUR+MED) et patients (PAT), on observe 3,525 épisodes de contact pour une durée cumulée de 166,320 s. Les 6 individus les plus connectés (IDs 1115, 1295, 1207, 1210, 1181, 1193 : tous NUR) concentrent 42.1% des contacts HCW–PAT et 44.3% de la durée cumulée correspondante. De plus, le nombre de contacts et la durée cumulée sont fortement corrélés (r = 0,979), et le nombre de patients distincts contactés est corrélé au nombre total de contacts (r = 0,694), ce qui est cohérent avec les observations de l'article.

3.4 Communauté et modularité

Dans le contexte d'un réseau de contacts hospitaliers, les communautés, c.à.d. les groupes plus densément connectés, peuvent refléter des groupes de travail ou des interactions fréquentes liées à l'organisation des soins.

Une méthode de maximisation de la modularité est ici utilisée afin d'identifier une partition du graphe agrégé en communautés. La visualisation du graphe agrégé, colorée selon les communautés détectées par la méthode de Louvain, met en évidence une structuration non aléatoire du réseau. Il faut cependant tenir compte que le graphe n'est pas hiérarchique et que sa représentation 2D peut déformer les distances et relations entre les noeuds. Les résultats en utilisant la méthode de Louvain donnent différents résultats selon la méthode de pondération : la pondération par durée cumulée montrent un nombre de communautés = 5 et modularité = 0,3670. Les tailles sont [19, 19, 18, 13, 6] (ordre quelconque). D'un autre côté, la pondération par nombre d'épisodes (Ncontacts) produit un nombre de communautés = 6 et une modularité = 0.3044. Les tailles sont alors [18, 17, 16, 15, 7, 2].

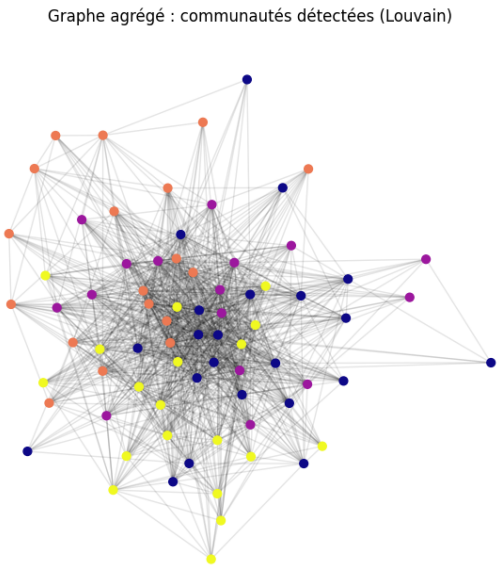

FIGURE 12 – représentation 2D des communautés d'après l'algorithme de Louvain.

3.5 Graphes pondérés et temporels

Jusqu'à présent, l'analyse s'est basée sur un graphe agrégé non pondéré, qui capture uniquement l'existence de contacts entre individus. Cette représentation simplifie l'analyse mais ignore l'intensité et la temporalité des interactions. Dans cette section, les poids sur les arêtes sont ajoutés, puis discutés.

Dans le graphe pondéré, nous considérons deux pondérations : durée cumulée, ou nombre d'épisodes (Ncontacts). Les centralités pondérées peuvent fournir une information supplémentaire en tenant compte de l'intensité des interactions.

L'article montre notamment que, malgré une forte variabilité temporelle des contacts, les mélanges (leur pattern) entre catégories d'individus restent statistiquement stables d'un jour à l'autre.

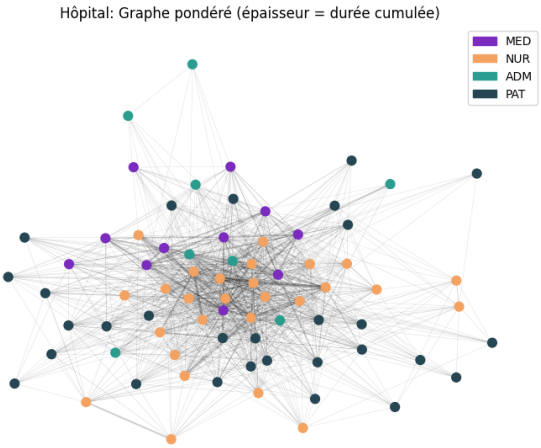

FIGURE 13 – Graphe pondéré sur la durée cumulée des contacts.

En reconstruisant les épisodes de contact selon la définition de l'article (c.f. section 1.2.), on obtient 14 037 épisodes de contacts et une durée cumulée totale de 648 480

secondes (environ 180.1 heures), en accord avec les ordres de grandeur rapportés dans l'article. L'utilisation de graphes agrégés peut conduire à surestimer certaines propriétés structurelles, telles que le clustering, et à ignorer des effets temporels pour les processus de propagation.

4 Comparaison et critique

L'analyse du réseau de contacts scolaires met en évidence une organisation fortement structurée par classes, avec une modularité élevée, traduisant une segmentation nette des interactions. Cette structure n'exclut pas une cohésion locale importante (triangles/-clustering au sein de certains groupes), mais elle limite les connexions entre classes : les liens inter-classes existent surtout via quelques individus, en particulier les enseignants. Les mesures de centralité confirment l'existence de noeuds critiques, notamment l'élève 1551 (3B), dont la position rend la diffusion potentiellement rapide, tandis que les enseignants jouent un rôle de ponts intercommunautaires. La comparaison à des modèles de référence (Erdős-Rényi, Barabási-Albert, Watts-Strogatz) souligne que les propriétés observées (forte modularité et organisation par groupes/rôles) sont difficiles à reproduire simultanément avec des modèles génériques. Enfin, les marches aléatoires et la simulation SIR suggèrent que, sans mécanisme de mitigation, la diffusion peut toucher une grande partie du réseau et n'est pas fortement dépendante du patient zéro.

L'analyse du réseau hospitalier, basée principalement sur des graphes agrégés, met en évidence un réseau plus petit mais très dense, avec un clustering élevé et une organisation fortement contrainte par les rôles (patients/soignants). La comparaison avec un null model de type configuration model montre que la densité de triangles observée ne s'explique pas uniquement par la distribution de degrés, ce qui suggère des mécanismes organisationnels (équipes, routines, interactions répétées). L'identification de "super spreaders" par le nombre et la durée des contacts est cohérente avec l'article, mais reste partielle : elle ne tient pas compte des chemins de transmission compatibles avec l'ordre temporel des contacts, ni de la position d'intermédiation (betweenness), qui peut signaler des points de passage entre sous-groupes. Une limite majeure est donc l'agrégation temporelle, susceptible de surestimer la connectivité effective pour la propagation.

En comparaison, le réseau scolaire est plus grand et structuré en communautés proches des classes, tandis que le réseau hospitalier est plus compact, très dense et dominé par la structure de rôles. Dans les deux cas, l'hétérogénéité des interactions implique que des interventions ciblées (sur des individus très connectés ou des ponts entre groupes) peuvent être plus pertinentes que des mesures uniformes. Une extension naturelle de ce travail consisterait à exploiter davantage la temporalité (chemins temporels, simulations SIR sur réseau temporel, analyses par journée) afin de mieux relier propriétés structurelles et dynamique de transmission.

Références

[1] Gemmetto, V., Barrat, A., Cattuto, C. *Mitigation of infectious disease at school : targeted class closure vs school closure*. BMC Infectious Diseases, 14 :695, 2014. doi :10.1186/s12879-014-0695-9.

[2] Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., et al. *Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors*. PLOS ONE, 8(9) :e73970, 2013. doi :10.1371/journal.pone.0073970.

[3] SocioPatterns collaboration. *SocioPatterns : high-resolution contact networks*. `http://www.sociopatterns.org`