HAL
open science

Epidemic processes on temporal networks

Anna Machens

To cite this version:

Anna Machens. Epidemic processes on temporal networks. Statistical Mechanics [cond-mat.stat-mech]. Aix Marseille Université, 2013. English. NNT : . tel-01287743

HAL Id: tel-01287743

https://hal.science/tel-01287743v1

Submitted on 14 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AIX-MARSEILLE UNIVERSITÉ

ÉCOLE DOCTORALE 352
Physique et Sciences de la matière

Thèse de Doctorat

présentée par

Anna Katharina Machens

pour obtenir le grade de

Docteur d'Aix-Marseille Université

Mention : Physique Théorique et Mathématique

À soutenir le 24 octobre 2013

Processus Épidémiques sur Réseaux Dynamiques

Directeur de thèse: Alain Barrat

Thèse préparée au Centre de Physique Théorique

Jury :

Rapporteurs :	Marc Barthélemy	-	CEA
	Renaud Lambiotte	-	Université de Namur
Directeur :	Alain Barrat	-	CNRS (CPT)
Co-Directeur :	Ciro Cattuto	-	ISI Fondation
Examinateurs :	Pablo Jensen	-	CNRS (IXXI)
	Jean-François Pinton	-	CNRS (ENS Lyon)

Acknowledgements

The research included in this dissertation could not have been performed if not for the assistance, patience, and support of many individuals. I would like to extend my gratitude first and foremost to my thesis advisor, Alain Barrat, for having given me the opportunity to work with the SocioPatterns Collaboration and do my thesis at Marseille. Without his encouragement and support this thesis would not have been possible. He has always been available and ready to answer any question. I am very grateful that he gave me the opportunity to present my work on international conferences and that he helped me extend my knowledge through discussions and by letting me visit les Houches and the Beg Rohu Summer School.

I am grateful to Marc Barthélemy and Renaud Lambiotte for giving me the honour to be my reviewers and to Ciro Cattuto, Pablo Jensen and Jean-François Pinton for having accepted to be members of the jury.

I would like to thank Ciro Cattuto for inviting me to ISI, for always being positive and for many fruitful discussions. It was always a pleasure to come to Turin, to have people to discuss my work with and also to learn what others are working on. I would like to extend my thanks to everyone at ISI for the warm welcome they gave me. I very much wish that some sort of collaboration continues or at least to visit again.

Thanks to the group of Carlo Rovelli for partly adopting me and to Sarah, Ben and Arnab for making the office a nice place to come to. At times, the office has been more of a home then my flat and so I can say, you were great flatmates.

All my gratitude goes to my parents for motivating and helping me through the ups and downs of writing my thesis. They were there for me at any time and never lost hope.

Synthèse en français

Introduction

Dans cette thèse, nous étudions l'influence des propriétés diverses des réseaux dynamiques et statiques sur la propagation des épidémies. Nous utilisons des données récentes de contacts en face à face de haute résolution. Pour les simulations de la propagation des épidémies, les nœuds du réseau sont divisés en compartiments de nœuds susceptibles (S), infectés (I), exposés (E) et guéris (R). Outre le modèle SEIR, des modèles avec moins de compartiments, comme SIR ou SI, seront utilisés aussi.

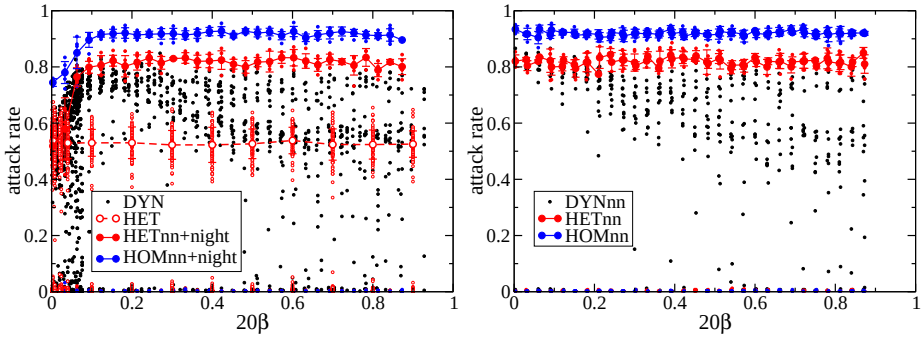
Récolte des données

Les données sont récoltées par la collaboration SocioPatterns. Ce sont des données des contacts en face à face entre les gens. Elles sont très détaillées avec une résolution temporelle de 20 s. On sait donc qui est en contact avec qui, quand et pour combien de temps. En contraste avec les données récoltées par des questionnaires, ici on n'a pas de biais causé par la mémoire des participants. Ces données sont donc très utiles pour simuler la propagation des épidémies. Néanmoins il peut y avoir des erreurs dues à une défaillance du matériel ou au comportement des participants avec les badges RFID. Il faut donc nettoyer les données, ce qui peut mener à un biais causé par les décisions prises. Aussi les contacts entre les gens sont enregistrés seulement si les gens sont vraiment en face à face à une distance de moins de 2 mètres. Les maladies comme la grippe , par contre, peuvent se propager aussi avec quelque probabilité si les gens sont à une plus grande distance. Ces chemins de propagation sont ignorées ici, ce qui peut limiter l'envergure de l'épidémie. De plus, il y a des gens qui étaient présents mais qui n'ont pas participé à l'expérience. Nous avons utilisé la méthode de bootstrapping pour simuler l'effet qu'un échantillon incomplet a sur le résultat de la propagation des épidémies. En regardant le nombre final de cas divisé par le nombre de participants, nous avons pu constater que l'effet de la réduction de l'échantillon est plus grand si le nombre de liens dans le réseau était déjà faible. Un autre problème est posé par l'enregistrement des contacts. Celui arrive de manière probabiliste. Pour garantir qu'un contact est vraiment enregistré, il faut donc attendre au moins 20 secondes. Ceci donne une limite à la résolution des données. Nous avons testé l'effet d'une limite de résolution. On peut voir qu'en baissant la résolution la longueur des contacts est surestimée, menant aussi à une surestimation de la taille de l'épidémie. Il se pose donc la question de savoir si les 20 secondes sont une résolution trop faible qui influence le résultat de la simulation. Nous ne pouvons pas tester l'effet d'augmenter la résolution. Au lieu de cela, nous avons testé un modèle très simple dans lequel un contact va être enregistré avec une probabilité proportionnelle à sa longueur (en unités de 20 secondes) dans des intervalles de plus en plus larges. Réduire la résolution jusqu'à une résolution de quelques minutes n'a pas changé le résultat, si en même temps la probabilité d'enregistrer des contacts (et de leur donner une longueur en multiples de l'intervalle minimal)

est proportionnelle à la longueur des contacts. On peut donc supposer que baisser la résolution de 20 secondes n'a pas un effet très fort puisque beaucoup de contacts plus courts ne vont pas avoir été enregistrés.

Propagation des épidémies sur réseaux dynamiques

Nous avons testé la dépendance de la propagation des épidémies dans les paramètres β et μ en changeant β et μ d'une façon qui laisse leur rapport β/μ constant. Quand β et μ sont augmentés, les épidémies finissent de plus en plus rapidement. Ils parcourent donc des parties de moins en moins courtes sur les données dynamiques. En même temps, l'effet des fluctuations des données sur le résultat est augmenté. En regardant l'évolution du nombre des infectés et du nombre des guéris dans le temps, nous avons pu voir, que ce nombre suit de plus en plus des fluctuations sur courte échelle pour les grandes valeurs des paramètres β et μ. Cet effet montre que pour β et μ grand, le choix du temps de début de l'épidémie peut jouer un rôle important sur le déroulement et surtout sur l'envergure de l'épidémie, en particulier si l'épidémie commence juste avant la tombée de la nuit. Alors que pour β et μ petits, l'évolution de l'épidémie était indépendant des fluctuations plus faibles et plus courts. Dans ce cas, le résultat de la simulation était le même si la simulation était effectuée sur le réseau dynamique ou sur un réseau statique (HET) construit en agrégeant le réseau. Les liens entre les nœuds recevaient alors un poids donné par la durée que ces deux nœuds ont été en contact divisé par le temps total des données.



(a) with nights (b) no nights

Figure 1: Le taux d'attaque d'une épidémie simulé sur les données de "sfhh" en fonction de β. Le rapport $\beta/\mu = 51.84$. Dans le plot à droit, les nuits sont exclus des données. Des simulations sur des réseaux HETnn, HOMnn sont ajoutées comme référence.

En augmentant β et μ les nuits jouent un rôle de plus en plus important. Jusqu'au point (β_{nn}), où le peak de l'épidémie est toujours avant la tombée de la nuit, la taille de l'épidémie augmente. A ce maximum et pour β plus grand, la taille de l'épidémie est donc la même que pour des épidémies simulées sur des données dynamiques dont les nuits étaient enlevées (DYNnn). Des simulations sur ce réseau dynamique (DYNnn) ne montraient pas beaucoup de changement avec β (en gardant β sur μ constant) pour des valeurs de β plus petites que β_{nn}. Nous avons construit un réseau statique (HETnn) qui n'était basé que sur le temps dans lequel il y avait des contacts dans les données dynamiques. Des simulations sur HETnn étaient les mêmes que les simulations sur DYNnn pour $\beta < \beta_{nn}$. Par contre, ils diffèrent fortement des simulations sur DYN pour ces paramètres. Si nous incluons des nuits, faisant un réseau bimodal (HETnn+nuits), qui est le réseau statique pendant les jours et qui ne montre aucun contact entre des nœuds pendant la nuit, alors les simulations sur DYN et sur (HETnn+nuits) sont très similaires pour les paramètres

$\beta < \beta_{nn}$. L'effet principal dans cet espace de paramètres en utilisant un réseau dynamique est donc la présence des nuits. En augmentant β au dessus β_{nn}, la taille de l'épidémie sur le réseau dynamique commence à baisser alors que sur les réseaux statiques elle reste la même. Des effets similaires ont été trouvés [] et le plupart du temps attribués au fait que la distribution de temps entre les contacts est large. Ici nous avons trouvé que si l'épidémie finit rapidement, seulement une partie du réseau dynamique est parcourue et seulement des nœuds qui étaient présents dans cette partie pouvaient être infectés. C'est cette limitation qui joue le rôle le plus fort. En créant un réseau statique de cette partie limitée du réseau dynamique et en simulant la propagation des épidémies là-dessus nous avons trouvé que la taille finale de l'épidémie est comparable à la taille obtenue en utilisant le réseau dynamique. Nous avons simplifié le modèle encore plus. En prenant le réseau dynamique calculé sur toutes les données mais en le limitant à contenir soit seulement les nœuds qui ont été présents pendant l'épidémie, soit le même nombre des nœuds que sur la partie parcourue par l'épidémie. Dans les deux cas, le résultat a encore été comparable à celui obtenu en simulant sur les données dynamiques.

De plus nous avons créé des réseaux dynamiques avec des modifications des propriétés diverses. Pour chaque réseau nous avons simulé la propagation des épidémies pour des valeurs diverses du paramètre β en laissant β sur μ fixe. Puis nous avons regardé le nombre moyen de nœuds et de liens qui sont activés pendant un temps spécifique. Dans un réseau (time shuffle) nous avons changé aléatoirement les temps de début de contacts entre les liens. Ce réseau montre les mêmes fluctuations du nombre de contacts sur le temps, mais les corrélations temporelles sont complètement anéanties. Cela menait à une augmentation fort du nombre de nœuds et liens qui sont actifs sur un temps précis. Aussi la taille de l'épidémie augmentait.

En concluant nous pouvons dire que le nombre de nœuds qui sont présents pendant l'épidémie joue un rôle crucial pour la mesure de la taille finale de l'épidémie.

Représentation des données

Les paramètres β et μ peuvent aussi jouer un rôle quand on essaie de réduire la complexité des données. Les réseaux dynamiques que nous avons sont très précis et détaillés. Mais souvent ce n'est pas nécessaire pour obtenir des résultats de simulations suffisamment exactes. Il peut y avoir des propriétés qui ne sont pas importantes pour la propagation des épidémies, des détails qui sont en trop. En éliminant des détails peu nécessaires des données, les caractéristiques importantes du réseau ressortent et les données deviennent plus générales.

Nous avons trouvé que pour réduire la résolution temporelle sans avoir d'effets négatifs causés par une surestimation des longueurs de contacts, il est suffisant d'attribuer aux liens des poids proportionnels à leur activité. Par contre, même en faisant cela, il reste une résolution minimale nécessaire qui est définie par le temps qu'un nœud reste infectieux. Dans le réseau dynamique, le nombre de nœuds qui peuvent être infectés par un nœud infectieux est limité au nombre de nœuds avec lesquels il a été en contact. Puisqu'en moyenne un nœud reste infectieux pendant $1/\mu$ secondes, cette durée est une bonne estimation pour la limite minimale de la résolution temporelle. En baissant la résolution au-delà de cette valeur, la taille finale de l'épidémie commence à croître. Si β est très grand, des erreurs sur cette limite deviennent plus importants et il peut être nécessaire d'avoir une résolution plus fine. Nous avons simulé la propagation pour trois sets de paramètres - β et μ petit, moyen et grand- sur des réseaux avec des résolutions différents entre 20 s et la longueur totale des données. Nous avons trouvé que la distribution du nombre final de cas ne change pas, tant que la résolution des données reste au-dessus de cette limite donnée par μ.

Il peut aussi être raisonnable de limiter la résolution des données du point de vue de la structure du réseau. Surtout quand il s'agit des simulations épidémiques de grande échelle ou

des vaccinations, des résultats au niveau de nœuds spécifiques ne sont pas nécessaires. Alors pour la simulation aussi il est possible que des informations sur les nœuds spécifiques sont superflues. Dans l'épidémiologie, des matrices de contacts sont souvent utilisées pour faire des prédictions sur des épidémies. Pour construire les matrices de contact, la population est représentée en groupes, par exemple en groupes d'âge. Les entrées de la matrice correspondent au temps moyen de contact entre les membre de chaque group. Si on utilise la matrice de contacts, les probabilités de contact entre les gens diffèrent selon leurs groupes, ce qui est une amélioration par rapport au mélange homogène où tous les nœuds sont en contact avec la même probabilité. De plus, ces matrices de contact sont aussi bien généralisables. Nous avons ici pris des données des contacts entre des gens dans un hôpital. L'avantage de ces données est que les participants sont divisés en groupes selon leur rôle dans l'hôpital. Il y a des infirmières, des auxiliaires, des docteurs, des patients et leurs parents, puisque c'était un hôpital pour enfants. Néanmoins d'autres choix sont possibles pour les groupes. Nous avons donc testé des choix de groupes différents. Puisque l'information des connections individuelles disparaît en faveur de l'information des connections de groupe, pour perdre le moins d'informations possible, il est favorable que les connections entre individus des deux groupes ont des poids similaires. La matrice d'adjacence peut donc être obtenu dans une forme proche d'une matrice formée de blocs uniformes. De tous les regroupements que nous avons essayé, les plus proches à une forme de bloc étaient le regroupement selon les rôles des participants et un regroupement selon leur degré. Le plus loin était le regroupement aléatoire. Pour le regroupement selon le degré et celui selon les rôles nous avons testé l'effet du nombre de groupes sur le résultat de la propagation des épidémies. Pour mesurer l'effet de la perte de l'information individuelle à cause du regroupement, nous mélangeons les liens entre deux groupes et les liens des nœuds d'un même group. Puis la propagation d'une épidémie est simulée avec le modèle SIR sur ces réseaux différents. Bien qu'avec une augmentation du nombre des groupes la perte de l'information individuelle diminue et le résultat devient légèrement plus précis, il est bien plus important de bien choisir les groupes au lieu d'augmenter leur nombre. La meilleure façon de choisir les groupes que nous avons trouvé reste le regroupement selon les rôles des participants. Nous utilisons donc ce regroupement pour tester des représentations différentes des données. Outre le réseau statique avec poids hétérogènes (HET) et la matrice de contact (CM), nous construisons un réseau statique avec poids homogènes (HOM), ce qui est une représentation qui garde l'information sur la structure du réseau en écartant toute l'information sur les poids et une représentation qui est une matrice de distributions de durées de contacts (CMD). Elle ressemble à CM, mais inclut des distributions des poids pour chaque group. Dans la représentation de données CMD, au lieu de mettre le même poids moyen sur tous les liens entre deux groupes, nous avons fait un fit sur la distribution des poids entre chaque deux groupes et nous tirons des poids de cette distribution en les attribuant arbitrairement aux liens entre les deux groupes. En comparant la distribution du nombre final de cas pour les représentations de données différentes, il devient clair que c'est surtout l'hétérogénéité des poids de liens qui joue un rôle important. Les simulations utilisant des représentations de données qui gardent l'information sur la distribution des poids sur les liens, comme HET et CMD, donnent des résultats similaires aux simulations sur les données exactes, alors que les simulations utilisant CM, HOM ou bien le cas d'un mélange homogène surestiment fortement le nombre final de cas. Le taux d'extinction de l'épidémie est également plus bas pour CM et HOM, comparé à HET, DYN et CMD. Avec les informations sur les rôles des participants nous pouvons aussi examiner la probabilité d'être infecté lors du déroulement d'une épidémie pour les membres de chaque group. On trouve que dans les groupes des auxiliaires et des infirmières, cette probabilité est plus élevée que dans les groupes de patients ou de parents. La probabilité d'être infecté pour les docteurs est entre ces deux extrêmes. La distribution du pourcentage des infectés dans chaque groupe est aussi similaire pour les simulations utilisant DYN, HET ou CMD. Si la matrice des contacts est utilisé,

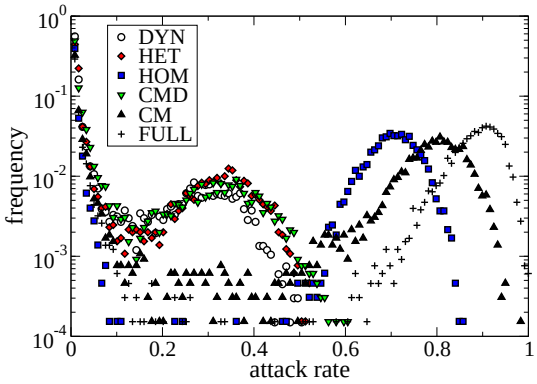


Figure 2: Distributions du taux d'attaque de l'épidémie pour les représentations de données différentes.

la probabilité d'être infecté est largement surestimée pour les patients et les parents. C'est dû au fait que entre ces deux groupes il y a très peu de contacts, mais avec une durée très longue, ce qui a pour conséquence que la moyenne est comparable à un cas avec beaucoup de contacts avec une durée moyennement faible. Seulement les représentations qui incluent l'information sur la distribution des poids peuvent distinguer entre ces deux cas. La matrice des distributions de contacts (CMD) est donc une représentation de données qui en même temps donne des résultats de simulations très similaires aux simulations sur les données exactes (DYN) et reste néanmoins généralisable. Dans le cas où les données des distributions de contacts sont accessibles, cette représentation est donc un meilleur choix que la matrice de contacts.

Immunisation

Bien que la nouvelle représentation des données fonctionne bien pour simuler l'envergure des épidémies, il est encore plus important d'être capable de les contenir effectivement. Nous testons donc combien de données sont nécessaires pour proposer des stratégies de vaccination utiles. La taille finale d'épidémie était différente pour les simulations sur diverses représentations de données, et nous constatons aussi que l'effet d'immunisation des nœuds change d'une représentation de données à une autre. Nous commençons avec des stratégies simples, immunisant 10 nœuds d'une même group. Nous comparons le nombre final de cas des vaccinations de chaque group. Si toutes les informations sont utilisées pour la simulation, il se trouve que le plus efficace est de vacciner les infirmières ou les auxiliaires et le moins efficace de vacciner les patients ou les parents. Le même ordre d'efficacité de vaccination de groupes est obtenu si nous utilisons des représentations de données avec hétérogénéité des poids (HET,CMD). Pour les simulations qui utilisent la matrice de contact par contre, il semble être aussi efficace de vacciner les patients ou des parents que de vacciner les infirmières ou les auxiliaires. Ceci mène donc à de fausses prédictions d'efficacité d'immunisation si la matrice de contact est utilisée pour simuler les épidémies.

De plus, car les informations dans les représentations de données sont plus ou moins limitées, des stratégies de vaccination qui peuvent être tirées de ces représentations différents sont plus ou moins sophistiquées. Pour chaque représentation (DYN,HET,HOM,CM,CMD) nous choisissons une stratégie de vaccination qui intègre le plus d'informations possible et qui nous semble optimale pour les données correspondantes. Nous testons l'efficacité de ces stratégies en simulant l'épidémie sur le réseau dynamique avec des nœuds vaccinés. Les stratégies obtenues sont

les suivantes: pour DYN et HET l'immunisation des nœuds par leur degré individuel, pour CMD l'immunisation des groupes par degré de groupe (et des nœuds dans les groupes sans ordre précis), pour CM l'immunisation des groupes par strength et pour le cas d'un mélange homogène l'immunisation arbitraire et sans ordre. Il se trouve que le plus d'information est inclus le mieux est la stratégie d'immunisation. Le désavantage de baser une recommandation d'immunisation sur des informations détaillées au niveau de chaque individu est le coût et la violation de la vie privée, mais aussi le manque de généralité. L'immunisation par degré de groupe donne des résultats assez bon par rapport à l'immunisation par degré d'individu mais sans ces désavantages. Cependant l'immunisation par strength moyen de groupe se révélait en partie même moins efficace que l'immunisation sans ordre.

Quand les données pour concevoir la stratégie d'immunisation sont limitées à la longueur d'une journée l'efficacité de la stratégie est réduite pour l'immunisation par degré individuel. Cependant, pour l'immunisation par degré moyen de groupe la réduction de l'efficacité est beaucoup moins forte. Les groupes des infirmières et auxiliaires restent les plus aptes pour l'immunisation même sur les données d'une journée seulement. Ces stratégies semblent donc être plus stables que les stratégies d'immunisation individuelle si la longueur des données est réduite.

Il reste à voir combien le résultat est influencé par le fait que la stratégie de l'immunisation est conçue sur les même données sur lesquelles il est testé. Dans un premier essai de tester cet effet, nous avons regardé le taux d'extinction de l'épidémie pour des épidémies très rapides. En prenant des parties de 24 heures pour concevoir la stratégie d'immunisation nous avons pu tester la dépendance du taux d'extinction en fonction du temps de début de l'épidémie. Il semblait que l'effet du temps de début ne dépendait pas énormément du tronçon de données sur lequel la stratégie était basée.

Jusqu'ici nous n'avons utilisé la même stratégie d'immunisation pour les données statiques (HET) et les données dynamiques (DYN). Il est possible que l'information dynamique puisse mener à des stratégies bien meilleures que la stratégie par degré individuel. Nous avons essayé de trouver une stratégie qui utilise les particularités des données dynamiques. Pour chaque nœud nous mesurons l'effet de son enlèvement en regardant tous les chemins temporels avec et sans ce nœud. Nous somme intéressés par le changement de la longueur des chemins temporels et le changement du temps des chemins entre deux nœuds. Nous définissons la 'signifiance' comme mesure qui accumule cet effet pour tous les chemins temporels qui commencent dans une fenêtre de temps limité. La distribution de cette 'signifiance' est exponentielle pour des temps de début différents, même la distribution des valeurs d'un seul nœud. Pour obtenir une stratégie d'immunisation nous prenons la moyenne sur toutes les valeurs d'un même nœud. En comparant cette stratégie avec des stratégies antérieures, elle n'est pas meilleure que la stratégie d'immunisation par degré individuel. Il est possible que beaucoup de l'information dynamique est perdu en moyennant la 'signifiance' sur les temps de débuts différents. De plus, pour simuler l'épidémie le modèle SIR avec $\beta < 1$ est utilisé alors que les chemins temporels sont basés sur un modèle SI avec $\beta = 1$. Comme nous avons vu auparavant, pour des épidémies plus lentes la structure à courte terme peut être négligée. Cependant, les structures à temps courts jouent un rôle très grand pour les chemins temporels.

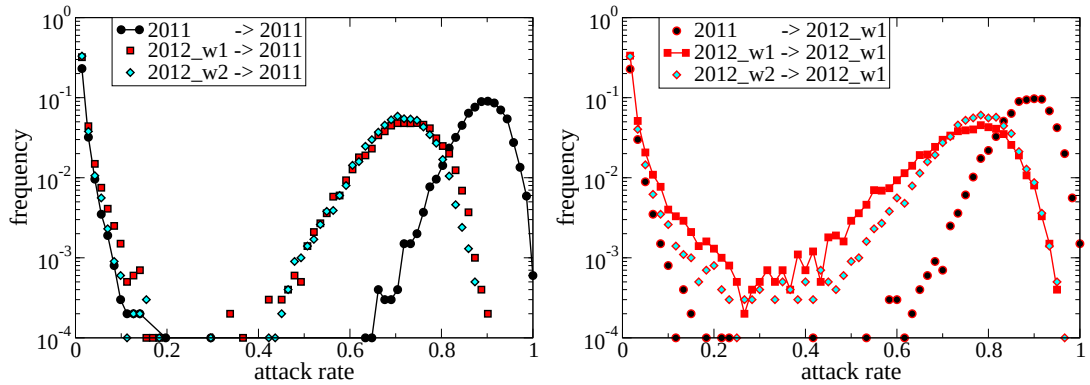
Prévisibilité

Pour qu'une représentation de données soit vraiment généralisable et qu'une stratégie d'immunisation soit efficace, il est important que les données ne changent pas trop. Si par exemple on crée une représentation de données d'un hôpital et on veut interpréter des simulations utilisant cette représentation comme valable pour un hôpital général et donc aussi pour n'importe quel

hôpital spécifique, alors les représentations de données obtenues à partir de différents hôpitaux spécifiques ne devraient pas être trop différentes. Une stratégie d'immunisation est conçue à partir de données réelles, mais doit être valable pour des situations futures. Pour pouvoir estimer les erreurs de prédictions il est donc essentiel qu'on puisse estimer les fluctuations des données.

Pour les données de l'hôpital à Lyon, "lyon2011" et "lyon2012", nous regardons les valeurs de la matrice de contact calculée pour chaque jour. Les valeurs sont assez stables la plupart du temps, c'est à dire que les fluctuations des valeurs ne sont pas beaucoup plus grandes que la différence entre les valeurs. Seulement la fin de la semaine dans les données de "lyon2012" marque une exception. Ici, les temps de contact moyens changent énormément entre la fin et le reste de la semaine. Ceci est clairement dû au fait que le travail est organisé différemment le weekend. Il y a par exemple seulement un docteur qui travaille et très peu d'infirmières. Par contre, une anomalie semble être la croissance du temps moyen de contact entre infirmières vers la fin des données de "lyon2012". Sans connaître les distributions du temps moyen des matrices journalières il est pourtant difficile de juger quelle matrice de contact est normale et laquelle est une exception dont la divergence des valeurs est causée par un autre effet.

Nous choisissons trois échantillons de 4 jours pour comparaison: les 4 jours des données de "lyon2011" et deux fois quatre jours des données de "lyon2012", dont quatre de la première semaine ("lyon2012_w1") et quatre de la deuxième semaine ("lyon2012_w2"). En calculant la matrice de contact sur quatre jours, les fluctuations sont encore diminuées un peu. Néanmoins, les matrices de contacts entre ces trois ensembles de données diffèrent toujours. Pendant que pour "lyon2011" le contact entre les médecins est le plus fort, dans les deux semaines de "lyon2012" c'est le contact entre les infirmières. La densité de liens entre des groupes montre les même tendances. La différence entre la matrice de contacts de "lyon2012_w1" et "lyon2012_w2" est seulement dans les valeurs exactes, par exemple, le temps moyen de contacts entre infirmières est beaucoup plus haut pour la deuxième semaine que pour la première, alors que la relation entre les valeurs est très similaire.



(a) prévision pour "lyon2011" (b) prévision pour "lyon2012_w1"

Figure 3: La taille final de l'épidémie, simulé sur la représentation de données CMD avec les paramètres $\beta = 100 * \mu$, $\gamma = 1/2\text{day}^{-1}$, $\mu = 1\text{day}^{-1}$. Dans la légende, "2011" $-$ > "2012_{w1}" signifie, que la simulation était fait avec la matrice de distributions CMD calculé sur les données de "lyon2011" pendant que les tailles de groupes pour construire le réseau étaient pris des données de "lyon2012_w1".

Nous comparons la simulation des épidémies sur des matrices de distribution de contacts (CMD) entre les trois ensembles de données. Alors que le pourcentage d'infectés est similaire

pour les deux semaines de "lyon2012", il est visiblement plus haut pour les données de "lyon2011". Dans les trois ensembles de données le nombre de participants dans chaque rôle n'est pas pareil. Un des avantages de la représentation de données CMD est que le nombre de personnes dans chaque groupe peut être choisi librement. Nous utilisons donc la représentation CMD pour "prédire" des épidémies dans des contextes différents, en prenant les paramètres de la distribution d'un ensemble de données et en l'appliquant à un autre ensemble de données, c.a.d. la simulation est effectuée avec les tailles de groupes d'un autre ensemble de données. En simulant les épidémies sur des réseaux créés de cette façon, nous voyons directement l'influence de la taille des groupes différents sur le nombre final de cas. De plus, changer le nombre de participants par rôle ne semble pas suffire pour améliorer les prédictions du nombre final de cas d'une épidémie. C'est surtout le temps moyen de tous les contacts qui diffère entre les trois ensembles de données. Alors que les simulations des deux semaines de "lyon2012" ne diffèrent pas beaucoup, la simulation utilisant les paramètres de la distribution de contacts de "lyon2011" ne donne pas des prédictions valables pour des épidémies éventuelles dans 2012. Si nous modifions les réseaux alors qu'ils n'ont pas seulement le vrai nombre de participants par groupe, mais aussi le bon temps moyen total de contacts, alors le nombre moyen de cas finals est très similaire à celui obtenu en utilisant la matrice des distributions de contacts construite avec les données de 2012. Néanmoins, le pourcentage de cas dans chaque groupe diffère encore puisque les valeurs relatives des matrices diffèrent encore.

Pour voir le développement des stratégies d'immunisation avec le temps, nous regardons le changement de l'ordre des degrés des nœuds sur un réseau dont la résolution temporelle est augmenté en agrégeant sur un temps court. Cet ordre change beaucoup. Un nœud qui a le plus haut degré à un temps spécifique ne l'a pas forcément dans le futur.

La distribution du degré d'un nœud pris à différent moments dans nôtres données, par contre, n'est pas aussi large que dans l'expérience de Braha [15]. Dans nos données, la distribution pour les nœuds avec le plus haut degré sur des fenêtres de quelques heures montre une crête. Le plupart du temps même les nœuds avec un degré très haut sur le réseau complètement agrégé (HET) ont un degré négligeable, mais quelques fois ils ont un degré très haut. Si les données sont agrégées sur des temps plus longues, ces fluctuations deviennent un peu plus faibles. En regardant comment l'ordre des nœuds change si on agrège de plus en plus longtemps, nous trouvons qu'il y a bien une limite de temps d'agrégation au-delà duquel l'ordre ne change plus beaucoup. Ce phénomène peut être dû au fait que nos données sont limitées, mais aussi au fait que dans l'hôpital les rôles des gens donnent naissance à un type de comportement spécifique. En tout cas, ici il suffit d'agréger les données pendant un temps assez court pour trouver les nœuds avec les plus hauts degrés sur le réseau HET. Au moins, déjà après très peu de temps, le degré moyen final (celui basé sur le réseau complètement agrégé) des premières 20 % des nœuds d'un classement selon leur degré réel (celui basé sur un réseau agrégé pendant un temps T) est très similaire au degré moyen final des premières 20% des nœuds d'un classement selon leur degré final. Néanmoins, le classement des nœuds continue à changer légèrement. En calculant le Kendall-τ du classement réel avec le classement final, on voit une croissance initiale forte qui devient rapidement moins raide. Pourtant, le Kendall-τ continue à croître avec T. L'ordre exact des nœuds continue à changer, surtout pour les nœuds de faible degré. L'ordre du degré moyen des rôles par contre est beaucoup plus stable. Très rapidement les groupes avec le plus haut degré moyen sont trouvés.

Distances

Il y a des propriétés ou des épidémies qui se propagent seulement pendant un nombre de pas limités. Des rumeurs peuvent changer à chaque pas de propagation. La distance entre deux

nœuds peut donc jouer un rôle important pour estimer si une épidémie va se propager d'un nœud à l'autre. Cependant, il est possible que la propagation d'une épidémie d'un nœud à un autre, qui est proche sur le réseau statique, prenne un temps long ou qu'elle soit impossible alors que l'épidémie se propage rapidement à un autre nœud beaucoup plus loin sur le réseau statique. La distance entre deux nœuds ne suffit donc pas pour estimer le temps de propagation. Ici nous étudions la longueur des chemins temporels. Un chemin temporel est le chemin le plus rapide entre deux nœuds n_1 et n_2 d'un réseau dynamique après l'introduction du nœud n_1 au temps t. Le chemin temporel de n_1 à n_2 peut être très différent du chemin temporel de n_2 à n_1. Bien que la longueur des chemins temporels soit corrélée avec la distance, elle est souvent beaucoup plus grande. Par exemple, deux nœuds qui ont une distance 4 sur le réseau statique peuvent être séparés de 8 pas si les contraintes temporelles sont respectées.

Comme la distribution des distances d'un réseau statique, la distribution des longueurs des chemins temporels peut caractériser le réseau dynamique. Nous comparons la distribution des longueurs de chemins temporels sur le réseau dynamique avec la distribution de longueurs de chemins d'un processus SI (chemins infectieux) sur un réseau statique. Si la dynamique du réseau temporel est poissonnienne avec une probabilité d'activité des liens très faible, les deux distributions sont les mêmes. Pour une activité forte des liens sur le réseau dynamique, la distribution des longueurs de chemins temporels change vers des longueurs plus basses jusqu'à devenir identique à la distribution des distances sur le réseau statique, si les liens sont actifs avec probabilité 1. Nous pouvons approximer la distribution de longueurs des chemins infectieux sur un réseau statique complètement connecté par la solution des équations différentielles du processus SI. Celles-ci donnent la distribution suivante des chemins infectieux:

$$p(d) = \frac{\ln(n-1)^d}{d!n}$$

avec une longueur moyenne des chemins infectieux de $\langle d \rangle \sim \ln(n)$, où n est le nombre de nœuds du réseau et d la longueur des chemins infectieux. Même pour des réseaux statiques complètement connectés cette distribution est seulement une approximation, parce qu'elle est conçue pour des réseaux de taille limite avec un nombre discret de nœuds. Si les réseaux ne sont pas complètement connectés, la distribution des chemins infectieux change. Nous testons comment la distributions des chemins infectieux dépend des densités de liens pour les réseaux Erdös-Rényi. La longueur moyenne des chemins infectieux reste stable pour une grande gamme de densités. La longueur moyenne ne croît rapidement qu'à l'approche du seuil de percolation. Nous regardons aussi l'influence du poids des liens sur la distribution des chemins infectieux. Dans ce but nous construisons des réseaux aléatoires dont les liens ont des poids tirés d'une distribution négative-binomial. L'influence des poids est seulement visible si le réseau est très dilué. Alors comparé à une distribution de poids étroite,une distribution de poids large augmente la longueur moyenne des chemins infectieux. Quant à la distribution des chemins temporels, le même effet est visible pour les réseaux avec une dynamique poissonienne, si la distribution des probabilités d'activité des liens est très large. Cet effet est beaucoup plus faible que l'effet de l'activité moyenne des liens. Finalement nous étudions la distribution des chemins temporels sur les données de l'hôpital à Lyon et les données d'une conférence. L'effet de l'activité simultanée des liens peut être observé pour les niveaux différents d'agrégation de données. Plus on baisse la résolution temporelle des données en les agrégeant, plus le nombre de liens actifs au même moment est grand et plus la longueur moyenne des chemins temporels est faible.

Pour voir l'effet de la dynamique des contacts, nous comparons la distribution des chemins temporels pour les données et des réseaux dynamiques basés sur les données mais avec une dynamique poissonienne (dHET et dHOM). Pour dHET et dHOM la probabilité d'activité d'un lien est donnée par le poids des liens dans les réseaux HET et HOM. Comme déjà vu pour

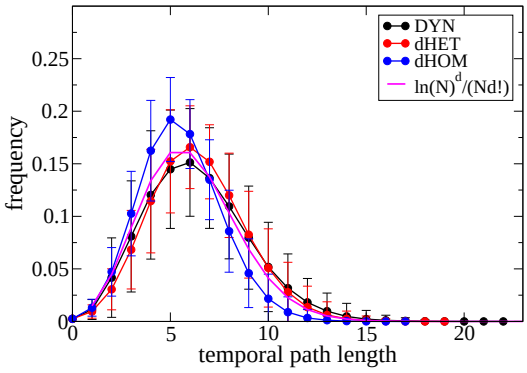


Figure 4: Distribution de la longueur de chemins temporels du réseau dynamique de contacts des données de "sfhh". La longueurs des chemins temporels des réseaux modèles dHEt et dHOM est montré aussi

l'influence de la hétérogénéité des poids, la distribution des chemins temporels a des valeurs moyennes plus grandes pour le réseau dHET que pour dHOM. La distribution de la longueur des chemins temporels pour le réseau DYN dépend des temps de début des chemins temporels. Nous avons pris la moyenne sur des temps de début différents. La dynamique des contacts ne semble pas avoir une grande influence sur la distribution de longueurs des chemins temporels. La distribution est très similaire pour les données et le réseau dHET. La longueur moyenne des chemins temporels n'est plus petite que la longueur moyenne sur dHET que pour un ensemble de données. L'explication de cette exception reste à éclaircir.

Conclusion

Dans cette thèse nous avons contribué à répondre aux questions sur les processus dynamiques sur réseaux temporels. En particulier, nous avons étudié l'influence des représentations des données sur les simulations des processus épidémiques, le niveau de détail nécessaire pour la représentation des données et sa dépendance des paramètres de la propagation de l'épidémie. Avec l'introduction de la matrice de distributions du temps de contacts nous espérons pouvoir améliorer dans le futur la précision des prédictions des épidémies et des stratégies d'immunisation en intégrant cette représentation des données aux modèles d'épidémies multi-échelles. De plus nous avons pu montrer comment les processus épidémiques dynamiques sont influencés par les propriétés temporelles des données. Il reste beaucoup de questions concernant la distribution des fluctuations des données dynamiques et son influence sur les prédictions des simulations des épidémies, les stratégies d'immunisation basées sur les données dynamiques ou la quantification des résultats avec des modèles.

Contents

1 Introduction **1**
1.1 Why networks? . . . 1
1.2 What are networks? . . . 1
1.3 How to classify networks? . . . 2
1.4 Dynamic processes on networks . . . 4
1.4.1 Epidemic spreading . . . 5
1.4.2 Simulation . . . 7
1.5 Network topology's influence on epidemic processes . . . 9
1.6 Overview . . . 10

2 Data collection **11**
2.1 SocioPatterns . . . 12
2.2 The datasets . . . 12
2.2.1 Activity . . . 12
2.2.2 Degree vs. Strength . . . 14
2.2.3 Contact-dynamics distributions . . . 16
2.3 Limitations of data . . . 17
2.4 A short note about cleaning . . . 18
2.5 Incomplete samples . . . 18
2.6 Discrete timesteps . . . 19

3 Epidemic simulation on temporal network data **25**
3.1 Activity fluctuations . . . 26
3.2 Influence of starting time . . . 27
3.3 Effect of nights . . . 29
3.4 Finite time . . . 31
3.5 Model networks . . . 34
3.6 Conclusion . . . 40

4 Data representation **41**
4.1 Time resolution . . . 42
4.2 Structural resolution . . . 44
4.2.1 Choice of groups . . . 48
4.2.2 Heterogeneity of weights . . . 52
4.2.3 Daily networks . . . 56
4.2.4 Influence of roles . . . 56
4.2.5 R_0-correction . . . 61
4.3 Conclusion . . . 64

5 Immunization on dynamic networks and data representations **65**
5.1 Influence of the data representations 66
5.2 Immunization strategies on static data representations 67
5.3 Effect of a limited time window 71
5.4 Time dependence of ranking efficiency 74
5.5 Immunization strategies on dynamic networks: significance 75
5.6 Conclusion 80

6 Predictability **81**
6.1 Degree ranking 81
6.2 Data-based predictions of epidemic spread 88
6.2.1 Comparing datasets 89
6.2.2 Effect of data variability on epidemic predictions 92
6.3 Conclusion 95

7 Distances **97**
7.1 Static distance vs. dynamic distance 98
7.2 Temporal path lengths and infection-path lengths 100
7.2.1 Discrete vs continuous 100
7.2.2 Influence of link density 104
7.2.3 Influence of the weight distribution 105
7.3 Distance on face-to-face contact networks 106
7.4 Conclusion 109

8 Conclusions **111**

A Appendix **115**

Chapter 1

Introduction

In this thesis, we investigate how various properties of temporal networks influence epidemic processes on networks. We are in particular interested in the role the data representation plays in this context and in how much detail of the data is necessary in order to obtain sufficiently accurate spreading results and decide on immunization strategies.

1.1 Why networks?

In order to tackle complex problems, simplification is a successful mechanism. By simplifying the problem as much as possible without losing its essential characteristics, it becomes better understandable. The language of network science does exactly this for complex systems. It describes the system as a set of interacting units. Only the connections between the often featureless units are important, all other properties of the system can be neglected. Systems simplified in this way can still lead to complex behaviour. Network theory has become one framework to study complex systems, thus supplying a common language to the study of topics as diverse as disease spreading, metabolic pathways or ecosystems.

Among others, human interaction patterns can now be quantified through dynamical networks of high resolution. This facilitates the study of transmission pathways and spreading of diseases, news or rumours between individuals. While in many previous studies of epidemic spread the problem has been simplified up to the point where interaction patterns were replaced by homogeneous mixing, describing the problem on dynamic networks enables a better understanding of the exact transmission pathways and the role of individuals for the epidemic spread. It also allows for a more accurate prediction of the epidemic outcome.

1.2 What are networks?

A network can be described as a graph $G(N, E)$, consisting of a set of vertices or nodes, N, and a set of edges or links, E.

Edges on the network can be physical or conceptual. In addition to the topology of the graph, we can store further information on the network, giving nodes and edges additional properties. Edges can have weights, describing, for example, the duration of contact between individuals, the number of contact events, the structural similarity between nodes, the correlation of events happening on the nodes, the trade volume between countries, or the flow of electricity, water or traffic. Edges can be one dimensional or multi-dimensional, describing different, independent

properties. In the latter case, we can represent them by multiplex networks. Nodes as well can have properties, like their infectiousness or resistance to epidemics, their capacity, political preferences or other characteristics. Again, these properties can be described by vectors as well. In addition to properties, nodes can have different states, which can change according to processes on the network. In the case of epidemic processes, the states can be, for example, susceptible, infected or recovered. The same is true for edges. If edges change their state depending on the process of the network, the network is adaptive. Nodes and edges can either be static and fixed in time, or their properties can vary with time. They can then be described as continuous functions $n(t)$ and $w(t)$. The weight function could for example represent the changing distance between two people over time.

In the networks used in this thesis, nodes do not have specific properties, but they can be in different states, which change according to the process on the network. Links between the nodes reflect the face-to-face contact between two individuals. The weight of the links is then either 1, if a contact exists, or 0 otherwise. Furthermore the weight-function is discrete in time. It is possible to integrate out the time component, taking the average over time of this function, which results in a static network with real-valued edge weights.

Often, a threshold is taken, setting edge weights below the threshold to zero. This threshold can be already present in the data collection or be applied afterwards. It can be a threshold in the affinity between two people, below which they are not considered friends, or a threshold in the distance between two people, above which they are not considered in contact. In networks in which edges are weighted by the correlation between nodes, a threshold can be introduced below which nodes are judged as uncorrelated and not connected by an edge. The choice of edges is essential as it defines the graph's topology [18].

If there are fewer edges than nodes, but the graph still forms one connected component, where all nodes can be reached from any other node via edges and intermediate nodes, the graph is called a tree. A (rooted) tree of N nodes has $N - 1$ edges and no loops. There is only one possible path between any two nodes. Removing any edge will make the graph fall apart into more than one connected component. If for two nodes, there is more than one path that connects them, the network contains cycles or loops. A cycle is independent of other cycles in the network, when it cannot be described by the sum of these cycles. The number of independent cycles in a connected network is $E - N + 1$ [12], so with every edge added to a connected network a new independent cycle is created.

The structure of the graph can be expressed in the adjacency matrix A_{ij}, where A_{ij} is one if nodes i and j are connected and zero otherwise. In the case of weighted networks, the matrix can also contain the weights w_{ij} of the edges between nodes i and j. For undirected graphs, the adjacency matrix is symmetric. The adjacency matrix can show a block-like structure, if the graph contains densely clustered subgraphs, and nodes in the same cluster are placed next to each other in the matrix columns.

1.3 How to classify networks?

As networks get bigger, the eye as a tool to understand the structure of the network is not sufficient anymore. The structure can become so complex that looking at it will not easily reveal the important properties of the network. Markers or proxies that will classify the properties of the network can be considered instead. Some of these classifiers, which were proven essential in different contexts, are described in the following.

- **degree**
  The degree k of a node indicates the number of neighbors that are directly connected to this

node via edges. It corresponds to the row sum of the adjacency matrix. The average degree of a network can be easily calculated as $< k >= 2E/N$. The degree distribution, which comprises information on the degree of all nodes of the graph, is a defining property of the network. Networks with a scale-free degree distribution are called scale-free networks in order to distinguish them, for example, from random networks (networks in which links are placed between nodes with some constant probability p), which have a Poissonian degree distribution.

- **strength**
  If the network is weighted, instead of the degree we can also define the strength of a node, which is the sum of the weights of all links starting from this node [10]

$$s_i = \sum_i w_{ij}$$

  If weights are placed randomly on all links, then the degree of a node is proportional to its strength $s \sim k < w >$ [10]. In the networks used here, nodes with high degree also tend to have links with high weight.

- **clustering coefficient**
  The clustering coefficient of a node n [107] is defined as the number of triangles $T(n)$ between this node and its neighbors, divided by the possible number of triangles $d(n) \cdot (d(n) - 1)/2$, where $d(n)$ is the degree of the node.

$$c_n = \frac{2T(n)}{d(n)(d(n) - 1)}$$

  If the clustering coefficient is high, then many of the neighbors of a node are connected among each other as well. The average clustering coefficient of a network is the average over all clustering coefficients of the nodes of the network. Social networks usually have higher clustering than random networks, as friends of the same node tend to be friends among each other as well. If the network is fully connected, the clustering coefficient of all nodes is equal to one. In a tree-network, none of the neighbors of a node are connected. The clustering coefficient is zero for every node.

- **path length**
  A path between two nodes n_1 and n_k is a sequence of nodes $P(n_1, n_2..., n_k)$ in which subsequent nodes are adjacent to each other. The shortest path in a network between node n_1 and node n_k is the path with the smallest number of adjacent nodes, over which information from node n_1 needs to pass to reach node n_k. Its length is called the distance between node n_1 and node n_k. The diameter of a network is the largest distance between two nodes on the graph. Graphs with a small diameter $D \sim \log N$ or a small average shortest path length have small-world properties [107]. This has been found to be the case for many social networks [59, 106].

- **betweenness**
  The betweenness centrality of a node n is related to the number of shortest paths d_{ij} between any two nodes i and j, which pass through node n. It is in fact the sum over all nodes i, j of the percentage of shortest paths $\frac{d_{ij}(n)}{d_{ij}}$ passing through n between any two nodes of the network [67, 32].

$$b_n = \sum_{ij} \frac{d_{ij}(n)}{d_{ij}}$$

It is a proxy for the influence a node n can have on messages passing on the network between any two nodes.

- **temporal patterns**
  Temporal networks can be regarded as a sequence of static networks. As the temporal development of the networks is too complex to follow directly, some proxies can also be put forward for temporal networks.

  - **activity**
    To get an impression of the variability over time, the **activity** of nodes and edges, the number of active nodes or edges, can be plotted as a function of time. This way, temporal patterns become easily visible.
  - **average number of active nodes and links**
    Even if the number of active nodes and links per timestep does not change, contacts can end and new contacts can be formed. In order to visualize this variability in the network, we cam aggregate over a certain time period and consider the number of distinct active nodes as a function of the aggregation time. In the networks considered here, the **average number of active nodes** increases with aggregation time. Wherever the derivative of this measure in respect to the time is positive, we have an introduction of new nodes into the network at that timescale. The same is valid for links when considering the **average number of active links**.
  - **time-varying centrality measures**
    For each instant of the temporal network, the above mentioned properties like degree or clustering coefficient can be calculated. However, the network snapshots are usually very sparse and consist of many disconnected components of two or three nodes. A more interesting measure is to partly aggregate the network and to consider these properties as a function of aggregation time and starting time of the aggregation.
  - **temporal distributions**
    Additionally, the **distributions of the contact times** and the times between contacts characterize the dynamics of the network. The duration of contacts is important for the transmission of diseases or information. If network characteristics are chosen in order to compare different networks in the face of their spreading capability, the **waiting-time distribution** [101] can also be of interest. The waiting time here is the time that passes between the start of two subsequent contacts of the same node with two diverse nodes. It characterizes the time that information which arrives at one node needs to wait before it can be passed on. The waiting-time plays a role when nodes can recover from the epidemic or stop transmitting information after a certain time. The **inter-contact time distribution** is here the distribution of the times in which single links are not active. It informs about the time that passes before a certain link is active again.

    In the time-varying networks of face-to-face contacts which are of interest here, the distribution of the duration of contacts as well as the distribution of waiting times are both rather broad [20].

1.4 Dynamic processes on networks

Nodes and edges can have different states. Networks represent the connections between different nodes. These connections can influence the states of the node. In order to understand how the

state of a node develops, influenced by the state of its neighbors, we study processes on the network. An introduction to complex dynamics on networks can be found in Barrat et al[9]. In the case of diffusion processes, the state of the node is characterized by the number or volume of diffusing particles which occupy the space of the node. Traffic flow can be modeled, for example, by diffusion processes on road networks or electricity flow on power grids. In these flow networks, edges can have a distinct capacity which cannot be exceeded. If there are no sinks or sources, the amount of particles or the volume of the flow is constant. In the case of spreading processes, there is no conservation of the entity (opinion, information or disease) which is spread. They can be used to model the spread of diseases or rumours, as well as the spreading or forming of opinions. In opinion spreading models like the voter model [19], a node adapts its state to the state of a random neighbor. In models of epidemic spreading, a node can transfer its state to a neighboring node if the neighbor is in the 'susceptible' state and the node itself in the 'infectious' state. Other state changes, like the transition from the 'infectious' to the 'recovered' state, happen independently of the influence of neighbors. Further processes with specific rules concerning the change of states of nodes can be used to better understand such diverse social phenomena as the consensus on word use in language (naming game [87, 27, 8]) or the evolution of cooperation strategies (prisoner's dilemma [69]). Here we will study more closely epidemic spreading processes on time-varying networks.

1.4.1 Epidemic spreading

To study epidemic spread in populations, the population is divided into different compartments. These compartments correspond to node states. People in one compartment are in the same state. We can distinguish between several compartments, for example, susceptible, infected and recovered. People change the compartment with a given transition rate. Depending on the names of the compartments, where S stands for susceptible, I for infectious, E for exposed and R for recovered, we have, among others, the SI, SIR and SEIR model. The different models are used in different situations. If people cannot recover from the disease and stay infected forever or at least longer than the time span which is considered, then the SI model is appropriate. The SIR model applies if people can recover, gaining life long immunity, immunity which is longer than the modeling time or if they die from the disease. A refinement of this model is the SEIR model, in which the time in which people carry the disease without yet being able to infect others is accounted for. Other models, where infected people become susceptible again after being infectious (SIS) or after having recovered (SIRS), are not treated here. We introduce the deterministic compartment models described by differential equations in a homogeneously mixing population [3, 47] and afterwards describe the simulation on networks for the SIR model.

SI model

The simplest epidemic model divides the population into two compartments, infected and susceptible. Each individual is in contact with all other individuals. Infected individuals (I) infect susceptible individuals (S) with rate β. The total number of individuals stays constant ($N = S + I$).

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta IS \tag{1.1}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta IS \tag{1.2}$$

The fraction of infectious and susceptible in the population are $i = I/N$, $s = S/N$. The differential equation which describes the spread of the SI epidemic over time can then be written

as:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta N i s \tag{1.3}$$

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta N i s \tag{1.4}$$

For this very basic model, an analytic solution can be obtained by replacing s in 1.4 with $1-i$ and i in 1.3 with $1-s$, and using a variable transform ($s = 1/y$, $i = 1/x$). It is

$$s(t) = \left(1 + e^{\beta N t}(\frac{1}{s_0} - 1)\right)^{-1} \tag{1.5}$$

$$i(t) = \left(1 + e^{-\beta N t}(\frac{1}{i_0} - 1)\right)^{-1} \tag{1.6}$$

The solution depends on the fraction of initially infected i_0 and initially susceptible s_0.

SIR model

In a slightly more realistic approach, a third compartment is introduced. Infected individuals can recover from the epidemic with rate μ. Recovered (R) individuals are removed. They are immune and cannot be infected again. The differential equations that correspond to this model are

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S I \tag{1.7}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta S I - \mu I \tag{1.8}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I \tag{1.9}$$

In contrast to the SI model, the whole population will not necessarily get infected in the SIR model. The fraction of the population which is reached by the epidemic depends on the parameters β and μ. If the propagation rate is very low compared to the recovery rate, it can happen that the epidemic does not spread at all. The basic reproduction number R_0 specifies the average number of secondary infections for an infected node in a population of susceptible individuals. The basic reproduction number depends directly on the rates β and μ as $R_0 = \frac{\beta}{\mu}N$. If the average number of secondary infections per node is below one, the epidemic will not reach a significant fraction of the population.

SEIR model

A further compartment that can be added is the compartment of exposed individuals (E). After being infected, individuals are not directly infectious. They enter first the exposed state in which they remain before becoming infectious with rate γ.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta S I \tag{1.10}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = +\beta S I - \gamma E \tag{1.11}$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = +\gamma E - \mu I \tag{1.12}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \mu I \tag{1.13}$$

However, in real life there is a finite discrete number of people, so that the epidemic spread needs to be modeled by difference equations. Propagation of the epidemic is not happening at fixed rates either, it is a stochastic process. Stochasticity can be integrated into the model using stochastic difference equations. Furthermore, not everybody is in contact with everybody else. Once the homogeneous mixing hypothesis is not considered sufficient and when the underlying network becomes too complex, for example when real data is given, differential equations become unpractical and unsolvable. However, the epidemic spread can still be simulated numerically on the underlying networks.

In addition to making the structure of the network more and more realistic, also the process on the network could be refined. Recovered individuals could become partly susceptible again. In general, different people can have different degrees of susceptibility to the epidemic, depending on their overall health condition or history. Some might even be partly immune. The propagation of the epidemic can furthermore depend on other variables, like seasonality or changing local conditions like temperature, humidity etc. We do not have any information on possible influencing conditions and also are mainly interested in how the network structure influences the spreading on the network, therefore we will limit the models used to simple compartment models. Parameters for the epidemic will be chosen arbitrarily, but within a reasonably realistic range. As the time scale on which we model is in the order of days, birth and death are not considered. However, due to the dynamic properties of the system, nodes are constantly introduced or removed from the system.

1.4.2 Simulation

We will describe the simulation for the SIR model here. Simulations for the SI and the SEIR model follow the same principles. We perform epidemic simulations on temporal networks based on contact data. The epidemic starts with only one infected node. Any effects due to the introduction of multiple infected seeds at various distances from each other on the network or at different times are thus excluded. Once the starting seed and starting time of the epidemic are chosen, the epidemic spreads stochastically with fixed rate β from infected nodes to neighboring susceptible nodes and infected nodes recover with rate μ. There are two different but related ways to simulate the spread of the epidemic. Either the propagation of the epidemic is advancing in discrete time steps, based on the discrete temporal network, or spreading is done continuously in time.

Choice of seed and starting time

In simulations, it is important how the starting time and starting node are chosen. Here we choose the node randomly with equal probability among all nodes. As nodes have different properties like degree, strength or presence, there are nodes which play a bigger part in the propagation of diseases. They are also more likely to contract the disease. The choice of the starting node introduces a bias, as not all starting nodes are equally likely in reality to be infected. As high degree nodes usually have a higher probability of being infected, one way of choosing the seed could be, for example, a function of its degree. However, since we do not know anything about the nodes of our dataset outside of the data, we do not use any information on the nodes in our choice of starting nodes. The seed is chosen at random. Similarly, we choose the starting time to be equally probable for all times. On static networks, the choice of a starting time is irrelevant. On dynamic networks, some thought has to be given to the choice of the random starting time. We consider two possibilities, a completely random starting time, independent of the choice of the seed node, or the time of first introduction of the seed into the network after a random point

in time. Since inter-contact times are bursty, so that nodes can be absent from the network for a long time, it makes a big difference if the starting time is chosen completely at random or as the time of first introduction of the node into the network after some random point in time. In the first case, the node has the chance to recover before entering the network, which will result in only one infected, in the second case, it will immediately be able to spread the epidemic. The two procedures make a difference for the extinction rate of the epidemic and thus also for the percentage of epidemic runs whose number of final cases is higher than a certain threshold. In comparison to static networks, infecting the seed only at its first introduction into the dynamic network would introduce a starting bias for dynamic networks. Seeds which are very rarely in contact with other nodes have small edge weights in the static network. Their potential for infecting their neighbors is low. If the seed is infected at the time of its first introduction into the network after some random time, its capacity for infecting neighbors will be much higher than on the comparable static network. Therefore, whenever the extinction rate is calculated or the number of simulation runs which result in a final size of the epidemic below a given threshold, we use the first case. If only the outcome of the epidemic above a given threshold is considered, both methods lead to the same result, but starting the epidemic with the introduction of the seed will need fewer simulation runs.

Discrete time steps

The minimum time step of the temporal network of the data sets we consider (see Ch. 2) is 20 seconds. Every 20 seconds the static network snapshots change. Many nodes will change their neighbors or become isolated. We therefore choose to use 20 seconds as the minimum spreading time step. Every 20 seconds, every infected node has once the possibility to infect each one of its current neighbors. The probability of infection during a contact of 20 seconds is $1 - \exp(-20\beta)$. Then, all infected nodes have the option to recover with probability $1 - \exp(-20\mu)$. For very small β and μ, these probabilities can be approximated by 20β for the infection of a neighboring node during the 20 seconds interval and 20μ for the recovery of an infected node. We use these approximations in all simulations with discrete time steps. When all infected nodes are recovered, the epidemic ends. The advantage of this method is that it is easy to implement, and computation time on the time varying networks which we use here is reasonably short. If the network has link weights w_{ij}, the probability of propagation will be modified to $20\beta' = 20\beta w_{ij}$. Link weights represent here the probability of activity of a link and will be always between 0 and 1.

Continuous time

Another option is to simulate in continuous time. On the static network, this simulation method is much faster. The probability for a node to become infected after t seconds of contact with an infected neighbor is then $-\beta \exp(-\beta t)$. For every infected node n the time of infection is stored. At the time of its infection, the time of its recovery is drawn from the exponential distribution function $\mu \exp(-\mu t)$. We do this by using the inverse probability transform sampling in order to transform random numbers x drawn from a uniform distribution into random numbers t for the exponential distribution. Thus the probability of the identically distributed random variable to fall into an interval x_1, x_2, $P(x_1 < X < x_2) = \int_{x_1}^{x_2} dx$, is set equal to the probability of the transformed random variable to fall into an interval $t_1 = t(x_1)$, $t_2 = t(x_2)$, $P(t_1 < T < t_2) = \int_{t_1}^{t_2} \int -\mu \exp(-\mu t)\, \mathrm{d}t$. Thus we have: $\int dx = \int \frac{\mathrm{d}x}{\mathrm{d}t}\, \mathrm{d}t = \int -\mu \exp(-\mu t)\, \mathrm{d}t$. The transformation function $t(x)$ is then the inverse of the cumulated probability function of the exponential distribution. The time of recovery is calculated as $t = -\ln(1 - x)/\mu$, where x

is a random number in the interval [0,1). For each neighbor of node n the time of possible transmission is drawn as $t = -\ln(1-x)/\beta$. If the time of transmission of the epidemic is earlier than the time of recovery of node n, then the neighbor node becomes infected at the given time. For all neighbors who can get infected by node n, the time of infection is stored. If they already possess a time of infection through another infected node, the time of infection is updated to the earlier of the two times. In some simulations, in which the exact infection path was of interest, also the node by which they got infected is stored. At the time of its recovery, node n recovers and is removed from the network. The epidemic is over when all infected nodes have recovered. If the network is weighted, for each transmission over a link with weight w_{ij} the probability of transmission β will be modified to $\beta' = \beta * w_{ij}$.

On the dynamic network, the simulation is identical except for one difference. The time of transmission is the time which passes while the infected node and its neighbor are in contact. In order to calculate the time of infection, the time in between the contacts of the two nodes has to be added, as long as the epidemic is not transmitted. The advantage of using this methods on the time varying network is that simulations can also be done for high β and μ. Furthermore, for each network snapshot of 20 seconds the epidemic can propagate along all connected paths and is not limited to infect only direct neighbors of the infectious node.

Basic reproduction number

We calculate the average number of secondary cases $\langle R_n \rangle$ for each node n in the network. To this end, we average the number of final cases that were directly infected by the starting node n over several epidemic simulations. In order to obtain an approximation for the basic reproduction number R_0 we then take the average over $\langle R_n \rangle$ for all nodes of the network. Again, this way of calculating R_0 is very dependent on the choice of starting time. When the epidemic is started with the introduction of the seed node, the extinction probability is much lower and thus the average number of secondary cases per node is higher. To calculate R_0 we therefore start all epidemics at random starting times, allowing nodes to recover before they come into contact. In any case, an R_0 calculated using the number of secondary infections of individual nodes is not necessarily equivalent to the epidemic threshold parameter calculated on population level models [16].

1.5 Network topology's influence on epidemic processes

When studying epidemic spread on static networks, it is of particular interest to understand how the topological structure of the network influences the course of the epidemic [46, 75, 49, 68, 64, 94, 81, 45]. To this end, often simulations on networks with particular structural properties are compared to simulations on random networks. It has thus been found, that the nature of the degree distribution has a strong influence on the epidemic threshold and the overall outcome of the epidemic. In particular, networks with a scale-free degree distribution facilitate epidemic outbreaks. In these scale-free networks the epidemic threshold is reduced, and epidemics propagate faster than in random networks [75, 76, 54, 65]. This has a direct consequence for immunization strategies. Random vaccination on scale-free networks is inefficient, as it often removes only nodes with limited importance for the spreading process, while targeted immunization can easily lead to a complete disintegration of the network by removing the hubs, nodes with high degree and consequently high importance [2, 23].

Another property of many social networks is a dense community structure [33]. While communities can facilitate spreading on a limited local scale, they globally hinder the diffusion of information. If on top of the community structure weights are correlated in such a way with

the topology that inter-community links have low weight and intra-community links have high weight [37], this effect of trapping information in communities is enhanced [70]. If community structures are strong, it is therefore a good immunization strategy to vaccinate individuals who bridge different communities [83]. Similarly, clustering slows down spreading [104] and can reduce the epidemic size [49] and R_0 [60].

Furthermore, when information on the temporal structure of the network is available, a comparison between simulations on dynamic networks with particular temporal properties and networks with randomized dynamics can inform on the influence of particular temporal properties of the network. In particular the burstiness of contact patterns slows down epidemic spread [44] while correlations between events can sometimes facilitate the propagation of epidemics [82].

1.6 Overview

The subsequent chapters treat the following subjects. Chapter 2 gives a short summary on the datasets used. Limitations of data and the effects of decisions concerning the data representations, such as the choice of the minimal time step length of temporal networks, are also discussed. In chapter 3, we investigate the effect of the dynamics of the network on the spreading process on the network. A focus will be put on the interplay between the timescales of the data and the process on the data, as well as the finite time effects of data. In the following chapter (Ch. 4) we follow two directions to simplify the data representation. On the one hand, we look at the optimal aggregation time of the temporal network, on the other hand, we try to simplify the aggregated networks by grouping nodes together. Here we introduce a contact matrix of distributions, which allows us to keep some of the heterogeneity of the links, even though single nodes adopt group properties, losing their individuality. We consider the efficacy of immunization schemes which can be derived from networks based on different data representations with different levels of detail in chapter 5. As we have a wide choice of data representations, we try to find a method which uses the maximum level of detail of the data in order to choose the optimal nodes for immunization, and we discuss its limitations. In chapter 6, we test for the amount of data necessary to make predictions for immunization schemes and the reliability of such predictions. We will also test the applicability of generalized data representations to other situations and discuss its limits. Finally, in chapter 7, we look at the relation of distances on static and dynamic networks and at the distribution of temporal distances as well as the distribution of the number of intermediary nodes in spreading processes. We give a short conclusion in chapter 8.

Chapter 2

Data collection

Network science has experienced a new surge of interest with the availability of large datasets. The discovery that many empirical networks describing systems relevant in diverse contexts share essential properties, like a scale-free or at least very broad distribution of degrees [22, 7], raised hopes that the theory of complex networks facilitates a general and unified theory of complex systems.

The study of the role of dynamical properties of temporal networks could only recently profit from detailed datasets. This opens the way for advanced, data-backed research on many open questions concerning the interaction of people, their dynamical contact patterns, the importance of single individuals on the spread of epidemics, temporal distances in networks and many more.

Dynamical (and topological) properties of real contact networks and their influence on dynamical processes on the network can be studied on the temporal data in order to understand which features of the dynamical properties have the most significant influence on specific dynamical processes on the network. These features can then be extracted, compared for different datasets and used in order to construct models of temporal networks [90, 78].

Data is also needed to inform existing models. For example, models for epidemic simulations can use census data, age data, data of human mobility and airline transportation data [99, 5]. Depending on the model, different degrees of precision of the data are needed. Precise data with high resolution is yet rare and therefore most models are using more general data. Comparing models informed by data representations containing various levels of detail can give an insight about how much information and which level of precision for data is needed in order to obtain results at a specific level of precision.

The data used here has a very high resolution. This allows us to compare results obtained by using precise data with results based on more general data - especially general static data - and discuss the differences due to various levels of detail on dynamic and individual information. The data was collected in the framework of the SocioPatterns collaboration [85], which we introduce in Sec.2.1.

Even though very precise data is a very good proxy for reality, there are still many limits. When a network is formed, decisions need to be taken about what is considered a node and what an edge. These decisions, like the choice of a threshold on the creation of edges, can greatly influence the topology of the network and the outcome of simulations. The right network representation for the given problem or question is therefore essential. [18]. Thus, in every data collection a bias is introduced through the choice of the network representation and selection of the information that is included (see sec. 2.3). Furthermore, datasets include errors, which can be eliminated by cleaning, but at the same time an expectation bias is introduced, as outliers

are discarded, when they do not meet the expected criteria of the corresponding distribution (see Sec. 2.4). In addition, datasets represent only a limited window of reality. We discuss the effects of its limitation in time in Sec. 3.4, its limitation in size in Sec. 2.5 and its limitation in resolution in Sec. 2.6.

2.1 SocioPatterns

The data sets used in this thesis were collected by the SocioPatterns collaboration [85]. They comprise face-to-face contact data between individuals at different venues. The data sets which we will use come from two settings: conferences and hospitals. Participants were equipped with radio-frequency identification (RFID) tags, which emit and receive signals in a peer-to-peer fashion. The emitted radio packets contain a unique identifier for the device and the time of emission. The tags register contacts autonomously whenever two participants face each other at a distance below 1-1.5 meters. The angle of detection is about 120 °. As radio signals are absorbed by the body water, the device can only efficiently emit signals towards the front of the body, thus greatly reducing the risk of false positive contacts of people who are in proximity but not facing each other.

The resolution of the contact data is extraordinarily precise, as contacts are registered continuously. However, detection of contacts is not instantaneous. The RFID tags alternate emit and receive cycles. When a packet is emitted during the emit cycle, it can only be registered by another RFID tag in the receiving cycle. Thus, it can take some time before a contact is registered. The contact data was therefore discretized into 20 second timesteps, which guaranteed with a probability of 99 % that an actual contact was recorded [20]. Also, meaningful contact lengths were assumed to not last much shorter than 20 seconds.

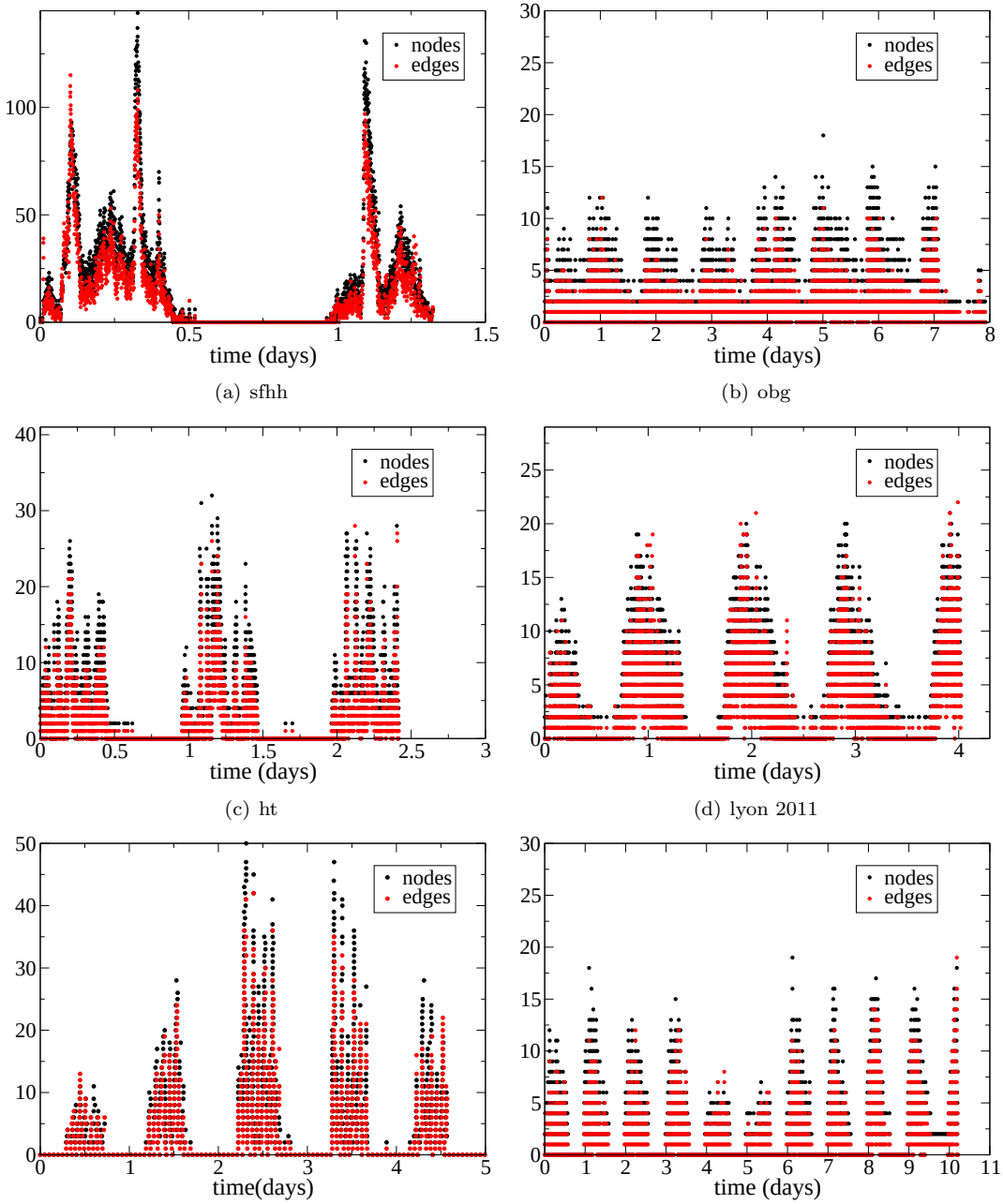
Further details about the collected data and the method of data collection can be found at the SocioPatterns website (www.sociopatterns.org) and in related papers [85, 20, 11].

2.2 The datasets

The data which are used throughout this work are face-to-face contact data of different venues, from the SocioPatterns collaboration [85]. The data vary in many properties, like number of participants, density of links or number of days. Depending on the setting, some data also have meta-information about the participants. For instance, the participants in the hospital data sets can be classified as Assistants, Doctors, Nurses, Patients or Caregivers. In order to understand the processes on the networks better, it is of advantage to know some of the structural and dynamical properties of the data. In the following, a short overview of some important aspects of the data are given. We use two types of data sets, data from conferences and data from hospitals. The conferences were the Congress of the 'Société Francaise d'Hygiène Hospitalière (sfhh), the European Semantic Web Conference (eswc) and the ACM Hypertext Conference (ht). The hospital data came from the Childrens' Hospital Ospedale Bambini Gesu (obg) in Rome and from a pediatric ward in a hospital in Lyon (lyon2011 and lyon2012). The length of the data set and the number of participants for each data set are given in Tab. 2.1.

2.2.1 Activity

In Fig. 2.1 the number of participants which are in contact with other participants is given for each timestep, as well as the number of connections which are active at each instant. This activity shows strong daily patterns. Coffee breaks and lunch breaks are marked by high peaks of activity,



(a) sfhh (b) obg

(c) ht (d) lyon 2011

(e) eswc (f) lyon 2012

Figure 2.1: Number of active nodes and links per timestep. On the left column are the conference data sets "sfhh", "ht" and "eswc", on the right column the hospital data sets "obg", "lyon2012" and "lyon2011"

Name	event type	year	length	participants	reference
sfhh	conference	2009	2 days	403	[88]
ht	conference	2009	3 days	113	[41, 84]
eswc	conference	2010	4 days	180	[1, 100]
obg	hospital	2010	10 days	119	[40]
lyon2011	hospital	2011	4 days	80	
lyon2012	hospital	2012	11 days	84	

Table 2.1: For each dataset, the type of the data (conference or hospital), the year in which it was collected, the time in days over which it was collected and the number of participants is given. References are provided for data sets which are publicly available or for which analysis have been published.

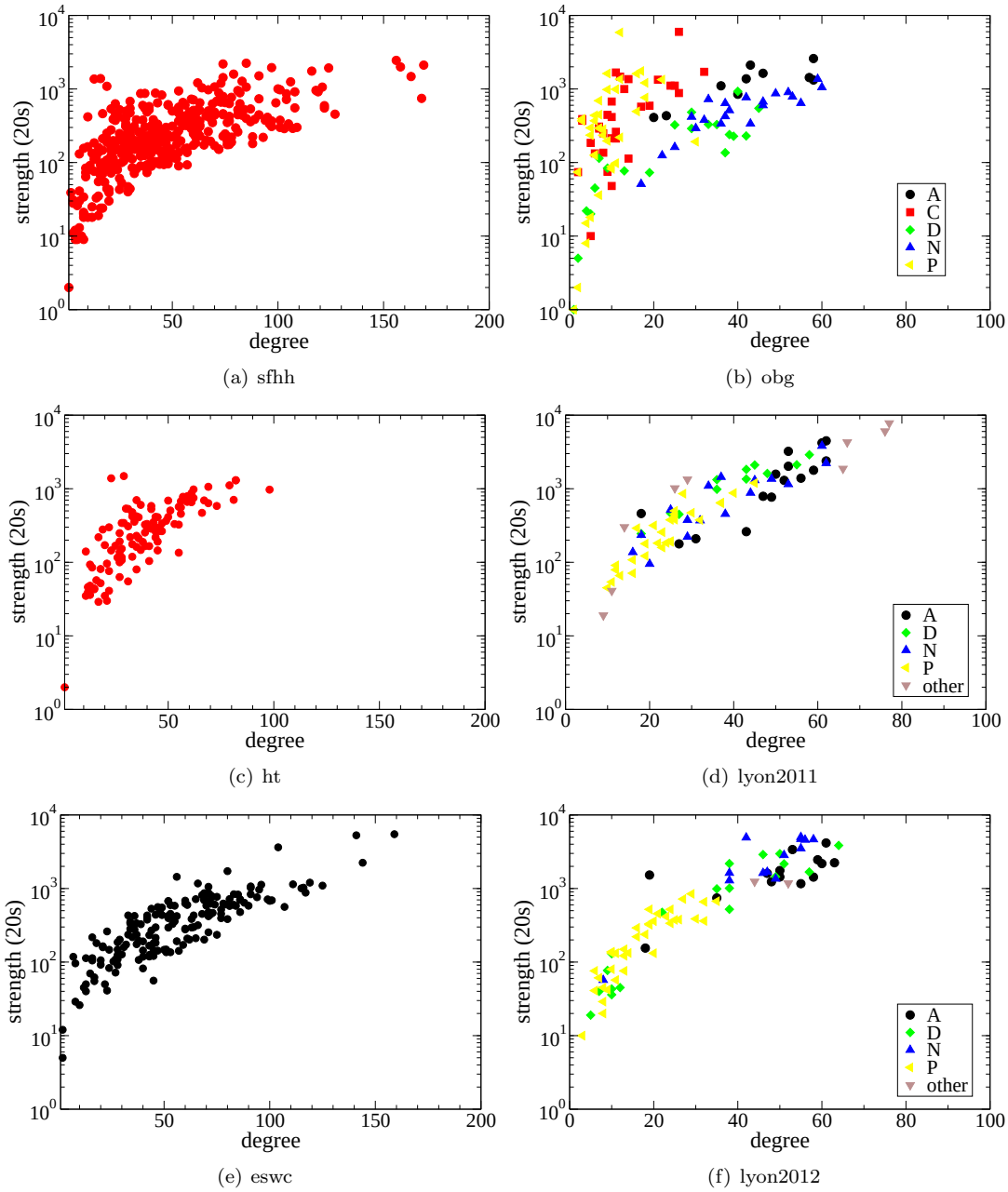
data	E	L	ld	W	$\langle w \rangle 10^{-3}$	$\langle d \rangle$	$\langle C \rangle$
sfhh	26040	9565	0.1181	1405220	1.284876	47.47	0.28
ht	9859	2196	0.3470	413540	0.901633	38.87	0.5347
eswc	18332	4890	0.3175	876920	0.415114	55.57	0.545
obg	16009	1227	0.1748	766000	0.911181	20.62	0.535
lyon2011	17672	1405	0.4446	885660	1.813264	35.13	0.6876
lyon2012	19015	1278	0.3666	962220	0.853234	30.43	0.74

Table 2.2: E: number of events L: number of links in the fully aggregated network, number of connections among nodes, ld: link density, W: total time of all contacts (in seconds), $\langle w \rangle$: average weight per link, corresponds to $W/(TL)$, where T is the length of the data set in seconds, $\langle d \rangle$: average degree, $\langle C \rangle$: average clustering coefficient

whereas during the night no contacts take place. Depending on the dataset, weekly variations can also be noticed. For instance, there is a strong drop in activity during the weekend in the "lyon2012" dataset and the first and last days of the "eswc" conference show lower activity. The percentage of nodes which are in contact at any time varies also between the datasets. While in the "obg" dataset, at no time more than about 15% of the participants are in contact at the same time, in the "sfhh" dataset at peak times it is over 30%. The networks also differ in the number of different contacts per person, the average number of events per time and other properties as listed in Tab. 2.2.

2.2.2 Degree vs. Strength

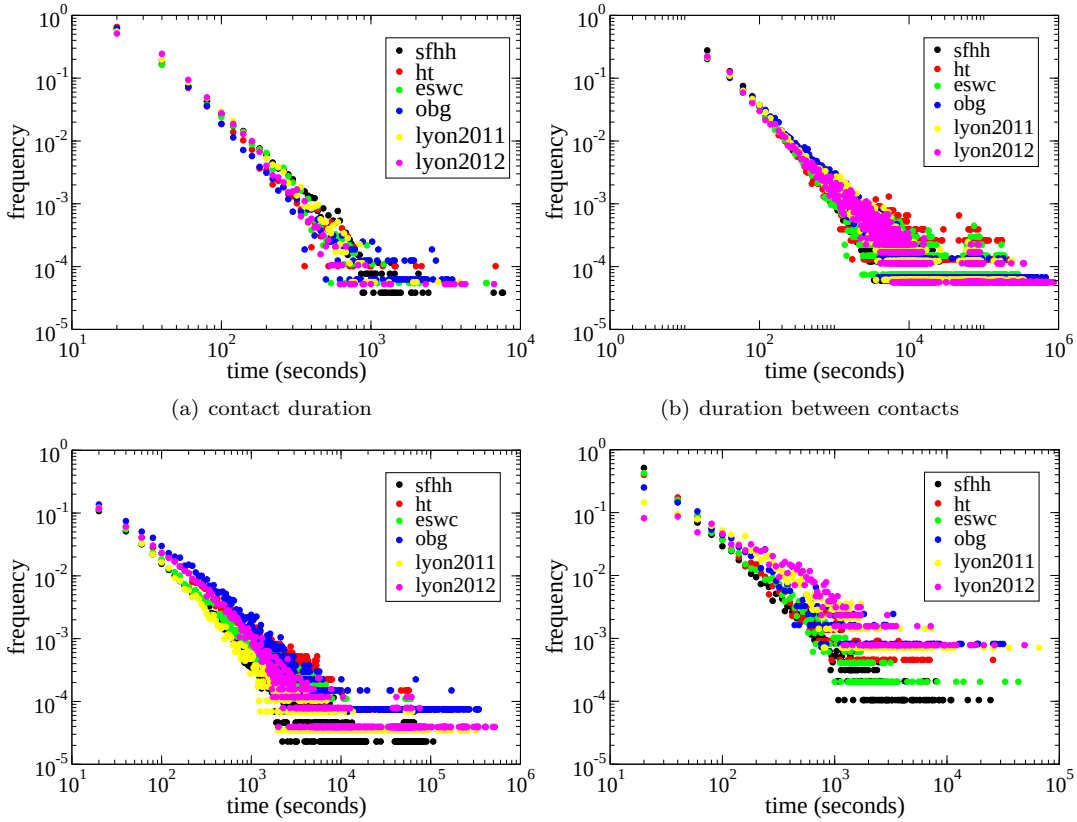
Nodes can be classified according to their degree and strength. If the contact time were identical for each link, then degree and strength would be linearly correlated. This is however not the case. As can be seen in Fig. 2.2, higher degree leads to higher strength, but not in a linear way. Participants with many diverse contacts also spend more time on average per contact. This superlinear dependence can be observed in many data sets and hints at the presence of superspreaders, individuals with many and intense contacts [20]. Furthermore, especially in the hospital data sets, the relation between degree and strength depends also on the class to which nodes belong. A particular case is the children's hospital ("obg"), where young patients were attended by their parents or other caregivers. For these classes, the relation between degree and strength is fundamentally different from other classes, as a low number of diverse contacts is set against high strength, resulting from few but very intense relations.



(a) sfhh (b) obg (c) ht (d) lyon2011 (e) eswc (f) lyon2012

Figure 2.2: For each data set (conferences at the left, hospitals at the right), the strength of the nodes is plotted over the degree. The nodes of the hospital data set have meta information concerning the role they occupy. The nodes of the "lyon2011" and "lyon2012" data sets can be differentiated into Assistants, Nurses, Doctors, Patients and others, where others are people from diverse roles. The "obg" data set includes Assistants, Nurses, Doctors, Patients and Caregivers. Patients were children in this hospital. Caregivers were parents or mentors which accompanied the child.

2.2.3 Contact-dynamics distributions



(a) contact duration

(b) duration between contacts

(c) waiting time

(d) total contact time

Figure 2.3: For all data sets are shown: (a) contact-time distribution, (b) inter-contact time distribution, (c) waiting-time distribution, (d) distribution of total contact time per link, which is proportional to the weight distribution, when divided by the respective length of the dataset.

The weight distribution is very broad. It is related to the distribution of total contact time per link. In Fig. 2.3(d) the distributions of total contact time per link for the six data sets are plotted. While the distributions for the conference data sets are very similar, the distributions for the hospital data sets diverge slightly for low contact times. Especially in the "lyon2012" dataset, there are comparatively few links with low contact times. This could be due to the fact that contacts in the hospital are less regulated by chance. Many contacts follow the schedule of the hospital, for example, when nurses or doctors visit the patients on a regular basis, or when they meet the same set of other nurses or assistants. In conferences short and unique contacts are more likely to happen by chance, as participants can mix freely during breaks. In the appendix, the weight distribution for contacts among and between different classes is shown for the "obg" dataset.

The contact-time distribution on the other hand is very similar for all data sets and shows behaviour similar to scale free distributions (see Fig. 2.3(a)). The distribution of the time that passes between two contact events among the same two neighbors also has a long tail, but a small peak appears at a duration of one day, or rather a small dip at about 12 hours. As contacts are

more likely to happen during the day, the duration between two contacts is less likely to be in the order of twelve hours and more likely to be one entire day, even though in general longer times between two contacts are less likely and short inter-contact times are abundant (see Fig. 2.3(b)). The waiting-time distribution (Fig. 2.3(c)) also shows broad behaviour with a peak due to the day-night patterns.

2.3 Limitations of data

The data collected via RFID tags has a great advantage over data collected by questionnaires. Contacts are registered directly and therefore not subjectively biased. This method also avoids some of the shortcomings of traditional data collection via questionnaires like the informant inaccuracy or the fixed choice effect [48]. However, other limitations cannot be eliminated. Datasets are a limited version of reality. They only show one instance of reality, limited in time, space and resolution. Datasets inevitably need to have a beginning and an end. Here, the time was limited to a few days. Also, only one specific place was surveyed, in this case a hospital or conference. As soon as participants leave the area, taking off their badges, contacts are not registered anymore. Participation of the experiments was on a voluntary basis and some individuals have declined to wear a badge. In case they also share behavioural properties, this could have introduced a sampling bias (see also Sec. 2.5). Some participants did not properly handle their RFID tags, which led to some spurious contacts (see Sec. 2.4). The data sets are also limited in time, extending only over a few days. For simulations which last longer than the dataset, an artificial repetition or extension of the data set is necessary.

Furthermore, data is always collected for a purpose, focusing only on those parts of reality which serve this purpose. Only one specific detail is considered and everything thought unnecessary is disregarded. In order to better understand communication patterns between people, many different datasets can be taken into consideration, each one adequate to answer a different question: telephone calls, email contacts, interaction in online forums, face-to-face contacts. To simulate the global spread of epidemics, census data, data on human mobility and travel data can be used.

A good proxy for the probability of transmission of diseases, like influenza, which are transmitted via aerosols or droplets in close contact [98, 38, 108], are face-to-face contacts. To be more specific, simulation of influenza transmission via small aerosol particles only requires information on room co-presence as particles stay air-born for some time, whereas larger particles settle quickly and require close contacts for transmission [38]. Nevertheless, other transmission pathways are possible as well. The transmission of mainly fomite-mediated diseases, for example, might not be simulated successfully by knowing only face-to-face contacts, as no information on touching or shared objects is registered.

Furthermore, a possible dependence of transmission probability on inter-personal distance might not be proportional to the distance-dependent detection probability of signals. The radio signals emitted by the RFID tags can have different predefined strengths, ranging from 1-2 meter up to several meters. The present data is registered with weak radio signals, limiting the data to true face-to-face contacts at a distance of 1-2 meters, which are a reasonable proxy for human communication. Influenza transmission can, however, also happen at larger distances through sneezing or small particle aerosols [53].

Other aspects which play a role in disease transmission, like information about the immunity of participants due to prior exposure, are not captured by the available data, mainly because they are unkown. Except for their contact activity, people are considered identical. Where meta-information is given, like information about the roles in the case of the hospital, which could be

correlated to frequent exposure and immunity, this information is not used to assign different properties to people. Another aspect of reality which cannot be captured in our simulations is the change of behaviour people undergo when faced with the possibility of infection vis-a-vis an infectious individual. During the data collection no individuals with respiratory illnesses were registered. However, all of these aspects can be considered as being part of the stochasticity of transmission and the choice of transmission parameters.

2.4 A short note about cleaning

Once the data have been stored, they need to be checked for errors and spurious contacts. Some participants do not handle the equipment as intended, taking off their badges or sometimes even leaving them next to other badges, which leads to continuously registered contacts. Some RFID tags could have been faulty, not detecting or emitting radio signals in a reliable way. Also, when badges are distributed at the beginning of the experiment or collected at the end contacts can erroneously be registered. These spurious contacts can be seen as abnormal behaviour in the data sets. High and narrow peaks of activity at the beginning or end of the data set, as well as outliers in the distributions of contact times, are an indication for spurious signals. Also, an augmented number of simultaneous contacts per node can be a sign of false positives. In some datasets, therefore, parts of the data were stripped off, some nodes were removed completely, and some partly if the number of simultaneous contacts exceeded a threshold of 6. In the "obg" dataset, only contacts which last less than one hour are considered.

However, deciding which registered behaviour is abnormal and due to a technical or human error, and which registered behaviour is real, is often difficult as no additional information is given to decide whether an outlier in the data is due to an uncommon event or due to an error in handling the material. Therefore, cleaning data will not only erase errors but also introduce a small expectation bias.

2.5 Incomplete samples

The observed interactions are just a subset of global interactions since people in the venue were not separated from the rest of the world. They could leave the venue, go home or to a hotel at night and meet other people with whom they could interact. Furthermore, participation was voluntary. Some people chose not to participate. Therefore, participants in the experiment were only a subset of present individuals. However, the fraction of people which agreed to participate was rather large in the used data sets. In the case of the "lyon2011" and "lyon2012" datasets, more than 90% of the individuals in the ward agreed to participate in the study. In the "obg" dataset, after cleaning, about 65% of the individuals are registered in the data. In the "sfhh" data, only about 30% of the conference participants agreed to wear RFID tags, in the "ht09" data it was about 75%. In addition, some contact events occurring during a measure might not have been registered. However, the probability of registering an event which lasted at least 20 seconds within a 20 seconds timestep was very high, so this does not play a major role here. Nevertheless, the detection of a face-to-face contact is subject to a threshold related to the distance of participants. Thus, many potential contact events taking place at a larger distance are not registered, thereby transforming the topology of the network [20].

The effects of incomplete data can be modeled through different sampling methods. An incomplete set of participants, for example, can be modeled by node sampling. If node sampling does not happen randomly, if for example people are more likely to participate in a study when

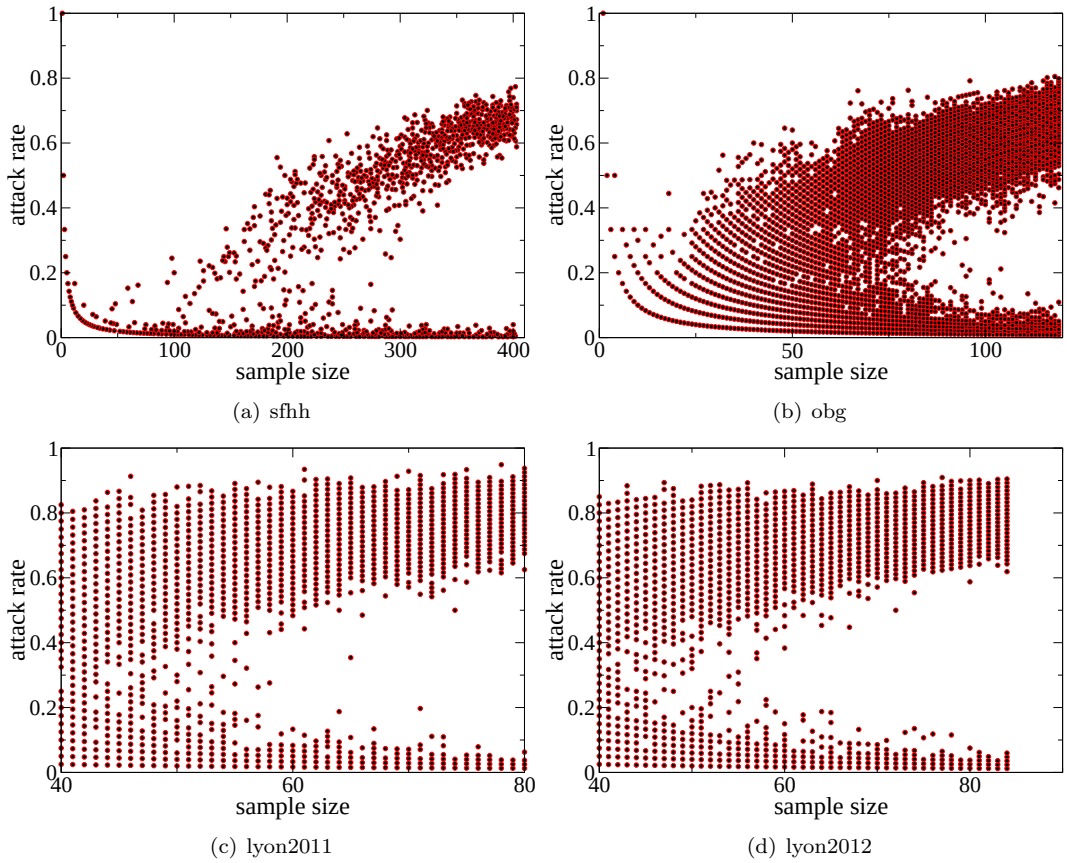
one of their friends participates, then people with more friends will be more likely to participate, introducing a sampling bias. This will inevitably change network properties like the degree distribution. Similarly it is possible that people are more likely to participate if a certain percentage of their community participates, which can have the effect that entire communities are excluded from the sample. But even random sampling does not leave the network unchanged. For example, a degree distribution which is scale free in the complete network might not be scale-free in the sampled network [92]. Other properties of the network change as well when the network is sampled [71, 50, 48, 26, 92, 91, 20]. In which way these properties change depends on the sampling method [50]. We will only consider incomplete data through node sampling. The related change of network properties has a direct effect on the outcome of processes on the network. The effects of incomplete data are inherently the same as the effects of node removal, for example through attack of nodes in a network [23, 24] or through immunization [77, 25].

We try to get an insight on the size of this effect on the outcome of epidemic processes on the network data. To this end we simulate an SIR model on random samples of different sizes. In Fig. 2.4 we see the results of sampling on the epidemic spreading for the "sfhh", the "obg" and the "lyon" datasets. The used sampling method is comparable to the vaccination of random individuals. As the distribution of the final number of cases is bimodal, removing nodes from the network has two effects, it decreases the average number of final cases for all runs that attain a certain percentage (here we generally choose 10%) of the network, and increases the number of runs, for which the outcome of the epidemic is below this percentage. For the "sfhh" network, the attack rate (AR), the final size of the epidemic divided by the number of nodes in the network, decreases already visibly with the random removal of very few nodes. Thus, in strongly sampled networks the fraction of the network that gets attained by the epidemic is underestimated. This effect, that the percentage of infected participants decreases with the removal of single nodes, is most likely stronger for sparse networks than for densely connected networks.

2.6 Discrete timesteps

The data used here is discretized into network snapshots at different time instants. The choice of the minimum timestep size proves to be quite important. We simulate the choice of different timestep sizes by aggregating the network with different aggregation timesteps. By doubling the aggregation timestep size, we merge two consecutive network snapshots into one. Whenever these two network snapshots contain the same nodes but with different links, then the new merged network snapshot contains nodes with higher degree. The chosen timestep length in the discretization of the dataset has therefore a direct influence on the distributions of the average degree of all nodes per timestep. In Fig. 2.5(a) the dependence of the average of this measure over all timesteps in the complete temporal graph on the length of the aggregation timesteps is shown. Longer aggregation increases the number of contacts which happen at the same time and thus the average instantaneous degree of the network at each timestep. In order to make results more consistent, nights were removed and data were stripped to a length which is a power of two.

When the network is aggregated into larger timesteps, the minimum connection time of contacts is also increased. For contacts which last for the length of one timestep, this means at least a doubling in contact length. Furthermore, several consecutive short contacts between the same two nodes can merge to a longer contact, and a large fraction of contacts is extended by one or even two timesteps, depending on their starting time and duration. The distribution of contact length depending on the aggregation timestep is shown in Fig. 2.6(a) for the "lyon2012" dataset. All contact lengths increase significantly when aggregation timesteps are doubled.



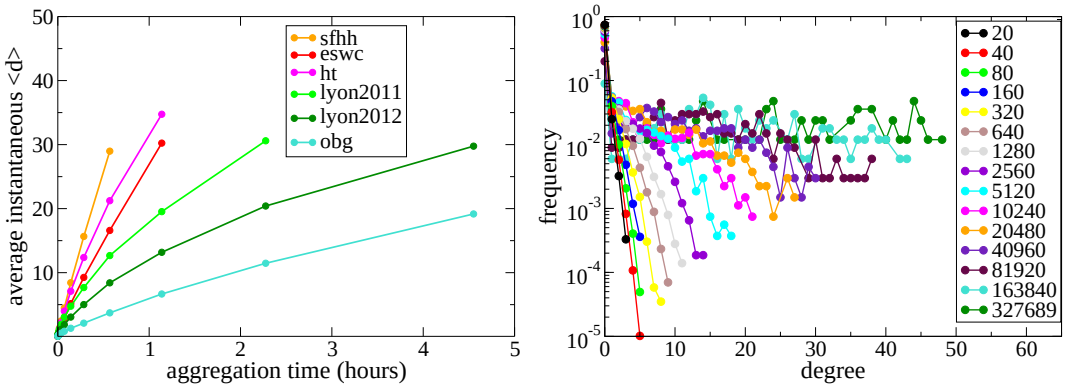
(a) sfhh

(b) obg

(c) lyon2011

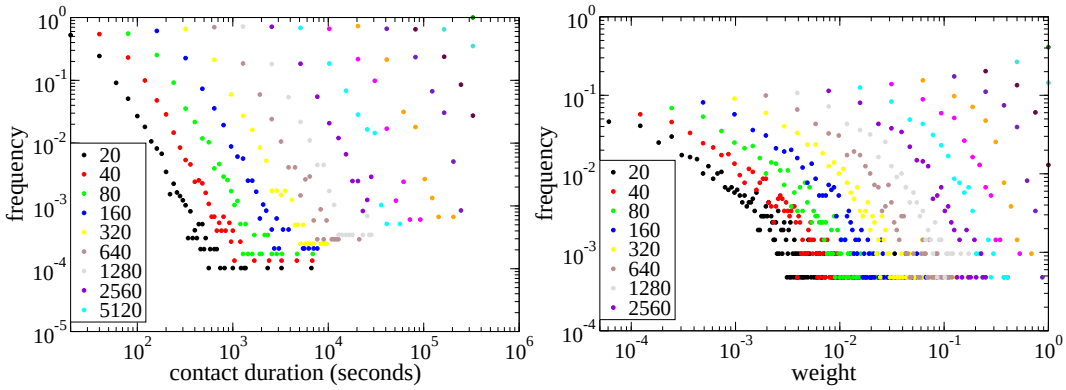
(d) lyon2012

Figure 2.4: The attack rate as a function of the number of participants in the data set. The parameters for the simulation are $\beta = 0.00015$ and (a) for the "sfhh" data set $\beta/\mu = 50$ (b) for the "obg" dataset $\beta/\mu = 500$, (c) for the "lyon2011" dataset $\beta/\mu = 100$ and (d) for the "lyon2012" dataset $\beta/\mu = 300$



(a) (b)

Figure 2.5: (a) The average degree of the network at each timestep, averaged over the complete network, dependent on the aggregation time step of the network for various temporal networks of different sizes and from different data sets. The networks were stripped of all contact-free episodes and shortend to a length which is a power of 2. (b) For the "lyon2012" dataset (with nights), the degree distributions are shown for different aggregation timesteps. The degree distribution is taken for each time instant of the dynamic network and then averaged over all degree distributions of the dynamic network with a specific resolution.

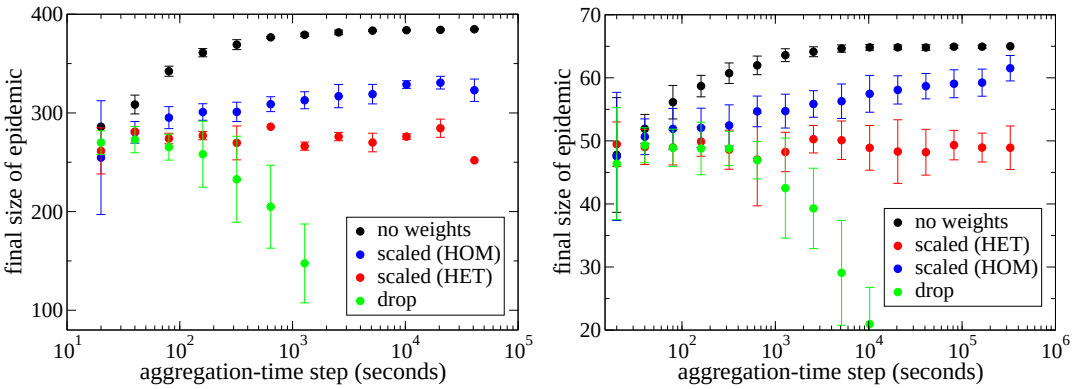


(a) contact duration (b) weight

Figure 2.6: (a) Contact-time distribution for the aggregated temporal networks with different aggregation time steps. (b) Distribution of weights for the aggregated temporal networks with different aggregation time steps

Related to the distribution of contact lengths is the total time a link between two nodes is active and the probability of activity for each link. When we build a static network out of the dynamic data, any contact that is registered in the temporal data corresponds to a link in the static network. Link weights of this resulting static network are chosen according to the total time each link has been active in the data divided by the length of the data-collection time. The distribution of the total time each link is active is then related to the distribution of weights. Epidemic simulations on this heterogeneous static network (HET) can be a good proxy for simulations on the temporal network [88].

The average weight of links plays an important role for the spreading of epidemics as it influences the probability of transmission. In Fig. 2.6(b) the distribution of weights on the HET network is shown for different aggregation steps of the network. With the increase of contact times for each event, also the weight of the links increases significantly. Aggregating over the complete network results in a static network where each link is active with probability 1. Therefore, choosing a higher minimum timestep size can have a strong influence on the epidemic process on the temporal network.



(a) sfhh (b) lyon2012

Figure 2.7: The average over the final size of the epidemic for epidemics which reach more than 10% of the population as a function of the aggregation timestep size for three different ways of aggregating the network. The parameters used are: μ^{-1}=4 days and (a) for "sfhh" $\beta = 50\mu$ (b) for "lyon2012" $\beta = 100\mu$

In Fig. 2.7 an epidemic is simulated on networks which were aggregated with this simple aggregation method. The influence of the minimal timestep size on the outcome of the epidemic is quite strong.

One solution to this problem is to introduce weights also on the links of the temporal network. When the network is aggregated, weights can give an alternative to distinct contact lengths. The weight here corresponds to the probability that a contact link is active. The exact times at which a link is active are thus replaced by activation probabilities. If weights are introduced even in the temporal network, the effective duration which links are active can be reduced below the minimum timestep length. Two alternative methods are chosen here which keep the average strength of the temporal graph constant. In the first method, each link has a weight which corresponds to the probability of activity for this link. This method also preserves the weight distribution of the links. While aggregating the network over n timesteps of size dt, the weight of a link between nodes i and j is defined as the total time T_{ij} the link was active during the n timesteps, divided by the length of the new timestep: $w_{ij} = T_{ij}/(ndt)$. If the network is

aggregated over the complete temporal data using this heterogeneous rescaling of link weights, the resulting network is the static network with heterogeneously weighted links (HET).

Another method is to weight links according to the overall average probability for links to be active during this timestep. The total time $T = \sum_{i>j} T_{ij}$ all links are in contact during the n timesteps is divided by the number of links L which are active and the length of the new aggregation timestep ndt: $\langle w \rangle = T/(Lndt)$. This method retains the average probability of activity of links in the network over a time period ndt. By aggregating over the entire network using this homogeneous rescaling of link weights, the result is a static network with homogeneous weights (HOM). The weights in the HOM network are the average over all weights in the HET network.

By including the information of how long links have been active during each time step, the epidemic spreading depends much less on the aggregation timesteps size. Similar to the distinction between spreading on the HOM and HET network [88], the more individual information is included in the link weights, the better the outcome. It seems therefore more important to focus on measuring the exact duration a link has spent in contact, rather than the exact time a contact starts. However, this is only valid within a certain range of parameters. Simulations in Fig. 2.7 are done with a probability of recovery $\mu = 1/(4\text{days})$ which is longer than the used data. As will be seen in Sec. 3.4 and Sec. 4.1 for larger μ the rescaling with weights is not enough to assure a reliable outcome of the epidemic for all aggregation step sizes, some information on the timing of links cannot be neglected for accurate simulation results.

In order to introduce weights on the contact links, information about the exact contact time length in a resolution lower than the resolution of the network timesteps needs to be available. Being able to increase the timestep length when using weighted links might therefore not allow for much improvement on the current data collection methods, as the exact information on contact length is still necessary, but it can be used in order to reduce the size of the data afterwards. This will be discussed more thoroughly in the scope of data representations in Ch. 4.

One way to obtain similar results as with the rescaling method with heterogeneous weights, but with less precision on the measuring method concerning the exact duration of the contacts, would be to make the probability of detection for a contact proportional to its activity during the time interval. This method is called "drop" in Fig. 2.7. We model this artificial scenario by using the weights of the links in the temporal network and creating a new temporal network where links are kept or dropped stochastically with a probability proportional to their weight for each aggregation timestep. Thus, if the link was active during the entire time ndt, the weight would be $w = 1$ and the link is active during the aggregated timestep of length ndt. If however only in m out of n time windows of size dt the link was active in the temporal network, then in the aggregated network the link is active in the time window of size ndt with probability m/n. As contact events are dropped, some connections between nodes which lasted only for a very short duration in the temporal network are lost entirely. This corresponds to setting a threshold in the aggregated static network, dropping links with low weight with a certain probability. This method shows that if detection of contacts is not always 100% sure, if furthermore it is proportional to the actual time spent in contact, then choosing slightly higher aggregation timesteps does not have the same severe effect on the outcome of the epidemic as the basic aggregation simulated above suggests. At least up to a minimum timestep of about 5 min the results are similar to the method with rescaled weights.

Thus, even though for an ideal data acquisition the timesteps should be infinitely small and the detection probability of a contact equal to one, a coarser method of contact detection can still yield reasonable data for epidemic simulations. If detection of an actual face-to-face contact has a constant probability per infinitesimal timestep, then the probability of detection of the contact before t seconds have passed follows a cumulative geometric distribution. The data is

discretized into 20s timesteps afterwards, so that, whenever a signal of a contact has fallen into one of the 20s time segments, the contact is marked in the data for this time segment. On the one hand, in order to have information about the true distribution of contact length, it is necessary that a contact that is active during an entire time segment is registered with a very high probability for this time segment so that long contacts, which last over multiple time segments, are not disrupted. On the other hand, if a contact lasts shorter than the time segment, then the probability of detecting a contact event should be proportional to the fraction of the time segment during which the contact is active, so that the total contact time of all contacts is conserved on average and not artificially prolonged. For the choice of a minimal time step, a compromise has to be made between the overestimation of activity by choosing the time segments too long and a loss of continuity in contacts by choosing it too short.

Chapter 3

Epidemic simulation on temporal network data

Due to a lack of temporal network data, most simulations of epidemic spread have until recently been performed on static networks. However, simulations on simple dynamical model networks have already shown the important influence of the dynamics of networks on the outcome of epidemic simulations. For example, varying the frequency with which nodes change their interaction partner has an effect on the final size of the epidemic. The faster nodes mix, the higher the final size of the epidemic, influencing R_0 and the epidemic threshold [102, 103].

The now available dynamical data open up new opportunities to study the temporal features of datasets and their influence on dynamical processes on the data set. Data sets come from a variety of sources, representing different aspects of human interactions, like communication via telephone or email, online interactions, human mobility, proximity or face-to-face contacts [85]. Despite their diversity, these datasets show many common patterns, among others large inter-event time distributions [6, 101]. Furthermore, events are often correlated, so that one action is the result of another (emails are forwarded or answered immediately, the news received in a telephone call is transmitted to friends) or two events have a high probability to happen at the same time (people meet and interact in groups, contacts are more likely to happen at a certain time of the day, i.e at lunch breaks, or during the week as opposed to the weekend). This and other factors result in the overall activity of the data showing daily patterns, night-day patterns, weekly patterns and patterns of higher time order.

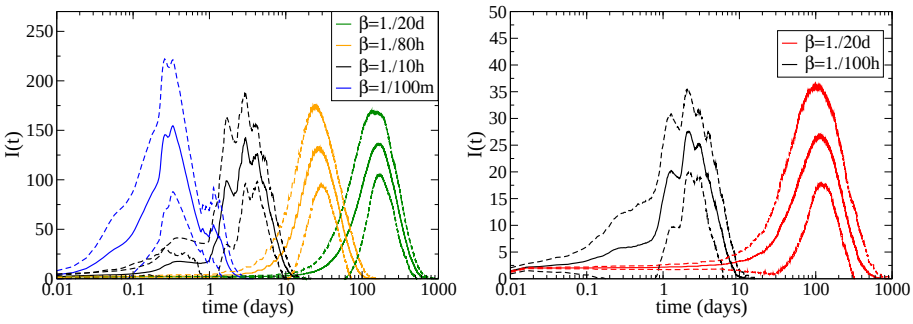
These dynamic structures of the data will influence dynamic processes on the data. In particular, the bursty nature of contact patterns was shown to slow down spreading [101, 39, 44], while temporal correlations can actually accelerate outbreaks [82]. Using the SIR model, it was found that burstiness hinders propagation for high propagation probabilities, while group conversations favor propagation at low propagation probabilities compared to temporal networks with randomized dynamics [63]. The heterogeneity of temporal contact patterns can even completely impede the spreading of epidemics [62].

In this chapter we have put our main focus on the timescales of the process and how they interact with the timescales of the network. For an SIR process on the temporal network, we can change the timescale of the process by changing the probability of propagation β and the probability of recovery μ at the same time. This is comparable to the mixing model of Volz et al [102] mentioned before. Increasing the speed of the epidemic is comparable to decreasing the speed of the temporal interactions, which corresponds to decreasing the mixing rate of nodes in the model of Volz. In temporal networks with large waiting-time distributions, speeding up the

epidemic will decrease the number of different neighbors each node will see during the course of the spread and especially during its infectious period, while slowing down the speed of the epidemic will lead to higher mixing of nodes during the epidemic. On static networks, changing the timescale of the process does not have any effect on the outcome of the epidemic. Only the duration of the epidemic process will change.

Dynamic networks have intrinsic activity fluctuations at different timescales. While the recovery of infectious individuals is unaffected by these activity fluctuations, the propagation of the epidemic depends strongly on the number of available neighbors and the overall density of the network. Temporal inhomogeneities in the network structure, changes of activity of nodes and links, can therefore temporarily change the equilibrium between recovery and propagation rates, thus steering the development of the epidemic. In section 3.1 we show how those temporal structures influence the course of the epidemic. In extreme cases, the outcome of the epidemic will then depend considerably on the starting time, as can be seen in Sec 3.2. Recurring long periods with no or very little activity can furthermore hinder the spread of epidemics, depending on how many of these periods are run through (see Sec. 3.3). In section 3.4, the influence of the timescale of the epidemic process on the size of the epidemic is shown for data sets with broad inter-contact time distributions. It becomes apparent that the length of the network dataset plays a non negligible role on the outcome of the epidemic. In order to compare which characteristics of the temporal network influence the outcome of the epidemic most, we simulate epidemics for different model networks in Sec 3.5, in which single features of the temporal network are changed.

3.1 Activity fluctuations



(a) sfhh (b) lyon 2011

Figure 3.1: Development of the number of infected nodes over time for different spreading parameters β and μ of the SIR process, where $\mu = \beta/100$. Out of 100 simulations, the continuous lines are the average over those simulations with a final size of the epidemic exceeding 10% of the nodes. The dotted lines show the standard error.

Temporal networks have dynamic structures at different timescales. These structures influence the dynamics of processes on the network. In return, processes on the network can reveal the temporal structure of the network. By tuning the timescale of the dynamic process, network structures at different timescales can become apparent in the development of the process.

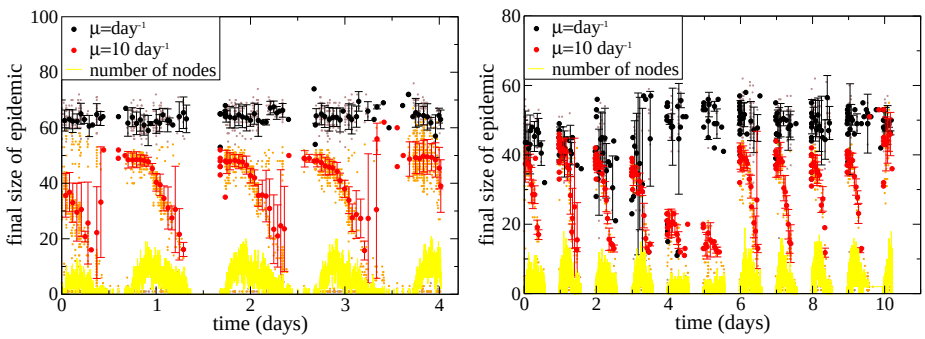
Fig. 3.1 shows how the number of infected individuals over time follows the activity pattern of the network for SIR processes at different timescales. Starting times are random. Processes which infect less than 10% of the nodes were filtered out before averaging. The average for different

random starting nodes and times is shown. The temporal evolution of the SIR processes was averaged relative to their starting time and not the absolute time on the network. The parameters β^{-1} and μ^{-1} are varied over several orders of magnitude from seconds to months, keeping the ratio β/μ constant. Thus β can be seen as an indicator for the spreading speed of the epidemic. For an epidemic spreading rate in the order of inverse minutes, daily patterns, like coffee breaks, characterize the course of the epidemic. For slightly slower spreading, with β^{-1} in the order of hours, night-day patterns are clearly visible. For even slower spreading (β^{-1} in the order of days), night-day patterns are still present but they become small in proportion to the overall shape of the time development of infected nodes. For β^{-1} in the order of months, the day-patterns disappear. The dataset used only extends over a few days; monthly or strong weekly variations are not present and can therefore not be seen in the course of the epidemic.

The number of newly infected nodes at any time is directly related to the probability of propagation and to the activity of the network. Thus, if the propagation probability is low, the same fluctuations in activity will result in low fluctuations in the number of newly infected individuals. When low activity leads to $\beta * \langle k \rangle / \mu < 1$, then the number of infected nodes decreases as more nodes recover than are newly infected. This threshold is an approximation on static networks when all nodes except the seed are susceptible. It can be applied for each static snapshot of the dynamic network where $\langle k \rangle$ is the average degree of the network snapshot but it is probably more useful to consider $\langle k \rangle$ as the average degree of the temporal network aggregated over the time a node is infectious and also include the weight of the links. The larger μ is, the steeper is the decrease of the number of infected individuals in periods without activity. Thus, the higher the probability for propagation and recovery, the more the variation in the data is imprinted on the spreading process.

For fast processes, patterns at small timescales are emphasized in the development of the process but the process also finishes quickly. A process with low β and μ covers a longer time period on the network but will also be less sensitive to short-term fluctuations.

3.2 Influence of starting time

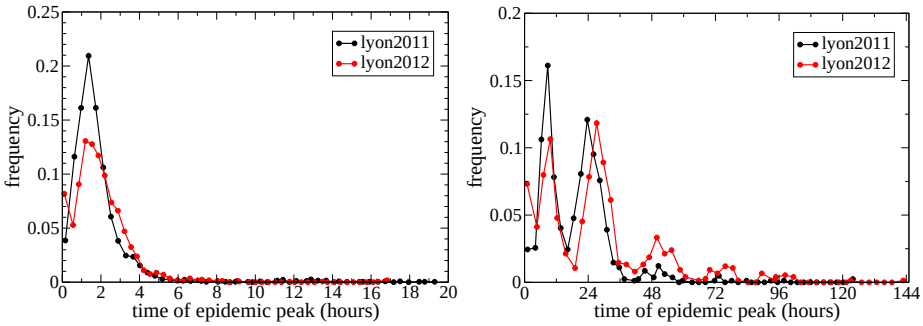


(a) lyon2011 (b) lyon2012

Figure 3.2: Final size of epidemic depending on the starting time of the epidemic for two different parameter sets with $\beta/\mu = 100$. A medium fast one with $\mu = 1\text{day}^{-1}$ and a fast one with $\mu = 10\text{day}^{-1}$. The node activity of the corresponding networks is plotted as a reference. The average and standard deviation is plotted over all outcomes with a final size greater 10% of the nodes and with a starting time falling into a one-hour bin. The grey and orange dots are the data points.

As the temporal data is not homogeneous in time, the starting time of a process on the network can greatly influence its outcome. In Fig. 3.2 it can be seen how the final size of the epidemic depends on the starting time of the epidemic. Depending on the speed of the epidemic, already a few hours difference can be decisive, independent of the intrinsic properties of the seed itself.

In the case of fast epidemic spread only a small part of the temporal network influences the epidemic process. In particular, if the process starts shortly before nightfall, the directly following night has a strong impact on the outcome of the epidemic. During the night, the ratio of newly infected to recovered nodes is much lower than during the day. The amount of infected nodes decreases, which either hinders the further epidemic spread or ends it altogether. In the case of slower epidemic spread the part of the network that is influential on the outcome of the epidemic can extend over several days. In Fig. 3.3 the distribution of the times up to the highest peak is shown. For fast epidemics, the epidemic peak is reached after only a short time. More precisely, for $\beta = 1./86.4s^{-1}$, the epidemic peak is reached around 1 hour after the start of the epidemic. In the slower epidemic, the probability of recovery $\mu = 1\text{day}^{-1}$ is lower, so that a significant part of infected nodes will not recover over night. Even though the epidemic peak is most likely to happen within the first day, the epidemic can also peak on the following days. In fact, it will have many daily peaks of different height. The peak time is calculated as the time with the highest number of infected individuals. The distribution of peak times (Fig. 3.3) shows daily recurrent peaks with diminishing intensity.



(a) $\mu = 10\text{day}^{-1}$ (b) $\mu = 1\text{day}^{-1}$

Figure 3.3: Distribution of the peak time of the epidemic for the "lyon2011"and the "lyon2012" dataset. The parameters of the epidemic process are $\mu = 10/day$ in the left and $\mu = 1/day$ in the right plot with $\beta/\mu = 100$.

However, not only the starting position relative to nightfall influences the epidemic. Less severe activity modulations, like the weekend in "lyon2012", also affect the outcome. For the very fast epidemic the effect is immediate. If the infected seed is introduced on the weekend, epidemics have an overall lower outcome as the activity during the beginning phase of the epidemic is generally lower. For the slightly slower epidemic, an introduction of the infected seed one or two days before the weekend leads to a smaller total number of infected, as the estimated peak time then falls together with the low activity on the weekend and the epidemic cannot expand as easily. Epidemics starting on the first day of the weekend will either end with a very low final size or outlast the low weekend activity and end with a high final size. Modulations in activity at higher timescales are not available here but will have an effect even in slow spreading epidemics.

3.3 Effect of nights

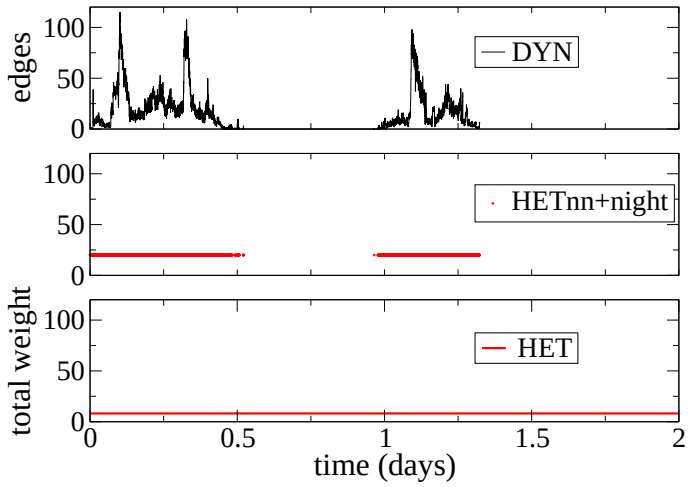
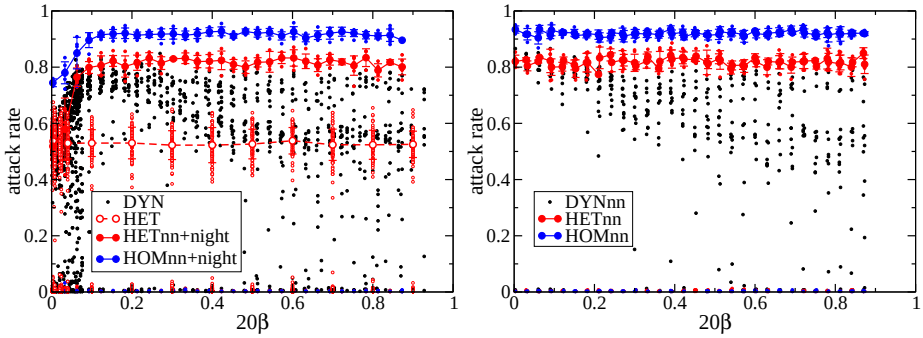


Figure 3.4: For the DYN network, the number of active edges at each time instant is shown. For HET and HETnn+night the total weight of all links, corresponding to the average number of active edges (at each time instant of the simulation), is plotted over time.

We will furthermore investigate the effect of the speed of the epidemic on its final size while the dynamic network's time scale stays fixed. In contrast to static networks, which do not have any time scales, temporal networks have a timescale defined by their activity fluctuation and contact patterns. Changing the timescale of the epidemic by varying β while leaving the ratio β/μ constant will lead to interactions between the timescale of the process and the network, affecting also the final size of the epidemic. As the largest variation in the course of the epidemic was due to night-day patterns, in Fig. 3.5 the dependence of the attack rate of the epidemic on the propagation parameter β for fixed ratio β/μ is plotted for the "sfhh" data set with and without nights. Since data sets are repeated for simulations, a second night was added to the "sfhh" data set, extending it to 48 hours. Thus, the nights are rather long, extending over 12 and 16 hours. The effect of the nights is accordingly quite strong, as the overall duration of nights is even longer than the active phase. Simulations on static networks are added as reference. The HET and HOM networks are calculated by aggregating over the entire temporal network. For the HET network, links are weighted according to their probability of appearance in the temporal network. The link weight is calculated as the total time the link was active divided by the total length of the temporal network. For the HOM network, all links have the same weight, given by the average link weight of the HET network. In the right plot, the temporal network does not include night times (DYNnn). The HETnn and HOMnn networks were calculated by aggregating over the temporal network without nights. In the left plot, nights were then added to these static networks at the same time at which nights occurred in the dynamic network so that at night times no propagation is possible.

For the simulation on networks where the nights were eliminated, the epidemic size does not decrease for lower propagation and recovery frequencies, while for networks with nights the effect of nights sets in for epidemics with $\beta \lesssim 0.01$, reducing the final size of the epidemic. For this region of epidemic speed, the main effect which regulates the outcome of the epidemic is the effect of nights. Further variations in contact patterns, as they exist between the HETnn

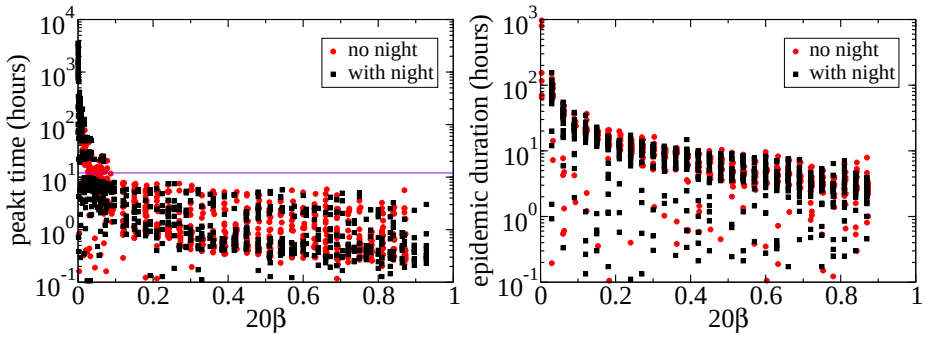


(a) with nights (b) no nights

Figure 3.5: The plot shows the attack rate of an epidemic (the percentage of final cases) on the "sfhh" data set as a function of the propagation probability β where the ratio $\beta/\mu = 51.84$ stays constant. The propagation probability β is increased from $1.5 * 10^{-4}$ to 0.05. In the left plot, the temporal network DYN includes nights. HETnn and HOMnn are calculated on the temporal network without nights, but periods of nights are added to these static graphs (see Fig. 3.4) so that HETnn+night and HOMnn+night also include periods of no activity during night times. The HET network is a static network, calculated by aggregating over the complete temporal network (DYN). In the right plot, the temporal network (DYNnn) does not include nights or time periods without any events. Simulations for the static networks HETnn and HOMnn which are based on DYNnn are plotted as comparison. For the static networks, the average over attack rates greater 0.1 for each parameter pair β and μ is plotted (big connected circles) in addition to the simulation outcome (HETnn -small red circles, HOMnn-small blue circles, HET-small red circles with white filling).

network with nights and the dynamic DYN network, are less important. The epidemic on the heterogeneous network HET which was calculated over the complete dataset including the nights reached about 50% of the nodes. This is the same as the outcome of a very slow epidemic on the HETnn+night network since for very slow epidemics the night-day patterns become irrelevant.

The faster the epidemic spreads on the network with nights, the fewer nights are run through and the higher the final size of the epidemic, up to the point where the epidemic ends before nightfall, and results on the temporal network with or without nights are identical. However, a second effect persists even in the temporal network without nights. The epidemic size also decreases for faster spreading epidemics.



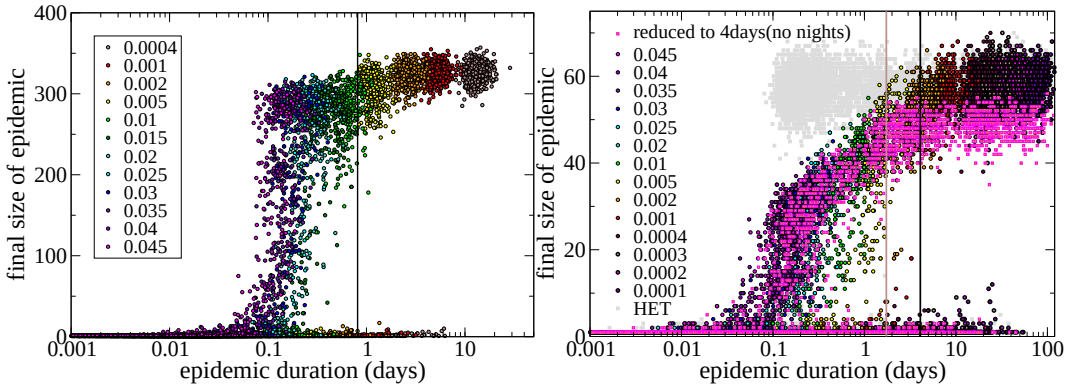
(a) peak time (b) epidemic duration

Figure 3.6: (a) Peak time of an epidemic and (b) epidemic duration with constant ratio of parameters $\beta/\mu = 51.84$. The propagation probability β is increased from $4 * 10^{-5}$ to 0.05. Simulations are shown for temporal networks ("sfhh") with nights (black) and without nights (red). The horizontal line marks half a day (12 hours).

In Fig. 3.6 the time up to the epidemic peak as well as the duration of the epidemic are plotted depending on β. For very slow epidemics the epidemic can also peak after having run through several nights, whereas faster epidemics do not extend over more than one day.

3.4 Finite time

The faster the epidemic ends, the less nodes are present during the course of the epidemic if node presence is not constant over the whole network but limited in time. In Fig. 3.7 the epidemic size is plotted against the epidemic duration. Simulations were done for various β, whereas for each β the epidemic has a relatively well defined span of probable durations. Nights are removed from the network in order to not mix different effects. The epidemic size increases with the epidemic duration until the duration has about the size of the temporal network. The size increases slightly further even when the network is run through once as nodes which are mainly present at the beginning of the epidemic could not have been reached yet. Any further increase of the duration linked to a smaller propagation probability β does not lead to a bigger epidemic size. To test the effects of the finite size, simulations are redone on a shorter version of the "lyon2012" dataset. Instead of the complete data, only 4 days of data are used. After removal of night times, this results in about 41.5 hours of data. Again it can be seen that the final size of the epidemic for durations much higher than the data length does not increase any more and is lower than the final size of simulations with the same parameters on the whole dataset. Whenever the epidemic lasts longer than the length of the temporal network, the final size of the epidemic might be



(a) sfhh (b) lyon2012

Figure 3.7: Epidemic size vs epidemic duration for various color coded pairs of β and μ, where $\beta/\mu = 100$ for the "lyon2012" network and $\beta/\mu = 50$ for the "sfhh" network. The value for β is marked in the legend. Both networks do not include night times. In (b) simulations on a reduced data set of 4 days (no nights) are added and marked with magenta squares for all β values. The grey line marks the length of the reduced data, the black line the length of the complete data. The grey squares show simulations on a static HET network as comparison.

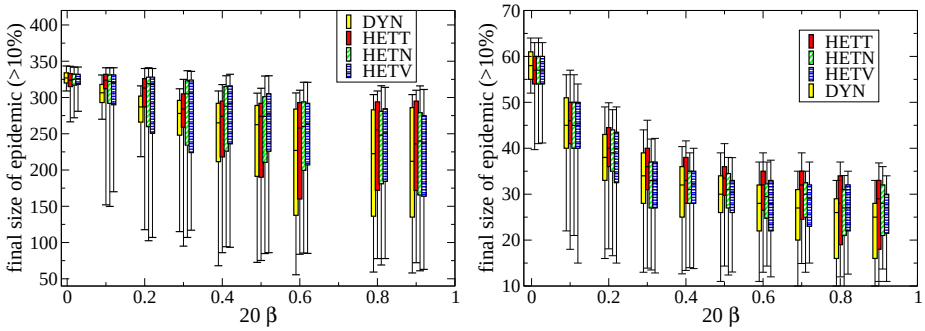
underestimated for simulations on the temporal network. This is most likely due to the fact that neither new nodes nor new links are introduced once the network is run through, as data is repeated without any change.

Thus, the main effect for the reduction of epidemic size is the fact that not all nodes could be reached if the epidemic duration was very short, as not all nodes had been present yet. This is at least the case here where the extension of the temporal networks leads to new nodes joining the network at later times and can be seen by comparing simulations on the complete data versus simulations on the network of 4 days.

For the same reason simulations on a static network, aggregated over the entire temporal network, overestimate the extent of the epidemic for high propagation probabilities and short epidemic duration. On the static network, all nodes and links are present all the time whereas on the dynamic network during the course of the epidemic neither all links nor all nodes could have been present. Spreading on a static network, however, gives similar results to spreading on a temporal network for epidemics which last about as long as the length of the temporal network. As can be seen as well in Fig. 3.7, the result for processes exceeding the length of the temporal network by far is then equivalent to the spreading on a static network with heterogeneous weights (HET).

Therefore, if an epidemic is simulated on a dynamic network on which nodes are continuously joining or leaving for extended time periods, the length of the network data is important. Repeating the dynamical network will not take possible new nodes into account. Artificially adding nodes or new links at later times is an option which however needs a theory of the dynamics. By changing or prolonging the original data based on hypotheses, much of the advantage of original, purely observational data is lost. On the other hand, epidemic simulations on aggregated static networks seem to give good results when the network is aggregated over a time slightly below the duration of the epidemic on the dynamic network.

We test this by comparing the spreading on the temporal network for various parameter sets with the spreading on various reduced static network. For each epidemic simulation on the



(a) sfhh (b) lyon2012

Figure 3.8: Epidemic size vs β for the DYN network and three static networks, HETT, HETN and HETV. For the "lyon2012" network: $\beta/\mu = 100$. For the "sfhh" network: $\beta/\mu = 50$. Both networks do not include nights. The boxes show the 25 and 75 percentiles as well as the median of the distribution of final cases, where only epidemics which reach more than 10% of the network are considered. The whiskers mark the 5 and 95 percentiles.

temporal network, simulations were redone on three different static networks, HETT, HETN and HETV.

- The **HETT** network corresponds to a static network which is aggregated over the exact length of the temporal network on which the epidemic took place. For each simulation run on the temporal network, we register the starting time and end time of the epidemic, aggregate the temporal network over this time period and start the spreading on the resulting static network with the same seed node.
- The **HETN** network is a subnetworks of the HET network, which is based on the entire temporal data set. For the HETN network, the HET network was reduced to include only those nodes which were present during the epidemic on the dynamic network.
- The **HETV** network is a subnetwork of the HET network. The HET network is reduced by randomly removing nodes until the number of nodes is identical to the number of nodes in the HETN network.

In Fig. 3.8 it can be seen that results averaged over simulations on these reduced static networks are similar to results on the temporal network for all parameter sets. Indeed, for faster epidemics with higher β and μ, the static networks used for simulations contain less nodes and less links. The main effect of the reduction of epidemic size for faster epidemics then seems to be the number of nodes and links which are present during the epidemic since just reducing the number of nodes on the static network will give similar results. Depending on the properties of the temporal network, the various methods to create reduced static networks can lead to a slight over or underestimation of the epidemic spread. In temporal networks in which nodes are preferably present over a certain finite time, nodes which are present during the epidemic are less likely to be present in the rest of the temporal data. Aggregating over the complete data instead of only over the duration of the epidemic will then add no new links to those nodes but reduce the weight of the already present links. Weights on links connected with these nodes are higher for the HETT network than for the HETN network. The HETN network will then have a lower outcome of the epidemic than the HETT network. If, on the other hand, node presence is rather homogeneous throughout the whole network, but new links are formed constantly, then aggregating over a

longer time will not dramatically reduce the weight of the links. However, it will add many new connections between nodes. Outcomes on the HETN network might then be higher than outcomes on the HETT network, as more links are added to the chosen nodes when aggregating over a longer time, and the weight reduction is not as strong.

If the importance of nodes in the network is very heterogeneous, then there are nodes which have many different connections and high activity in the network while others are only rarely present. Nodes which are important for epidemic spreading due to their high strength and high degree are also more likely to be present at any time of the temporal network. Therefore they are also more likely to be present during an epidemic. These nodes are then also more likely to be part of the HETN and the HETT network, whereas the choice of nodes for the HETV network is completely random. The choice of nodes in the HETN network is therefore slightly biased towards nodes with higher activity. Spreading on the HETV network might then have lower outcomes than on the HETN network.

Here, slight tendencies can be seen that nodes in the "lyon2012" network are not present over the entire dataset while nodes on the "sfhh" network have a higher probability of presence. In both networks, new links are introduced at all timescales (see also Fig. 3.10). However, these differences are not strong enough in order to significantly influence the simulation results. Results on the "sfhh" network are nevertheless slightly different. The maximum number of final cases in the distribution of final cases hardly decreases with β. This could be related to the high activity fluctuations in the "sfhh" data. In the next section, we form several model networks in order to better distinguish the properties of the data that are responsible for the reduction of the number of final cases.

In conclusion, it can be said that epidemic simulations on static networks can lead to results comparable to simulations on temporal networks if the aggregation length of the static network does not exceed the length of the epidemic. In cases where the epidemic duration is shorter than the time over which the network was aggregated, a reduction in network size can lead to better results if it is known how many nodes on average are present over a certain time. The main cause for the very different results on static and dynamic networks for fast spreading epidemics seems to be the number of present nodes and links over time, which is related to the distribution of inter-contact times and correlation between events. If the length of the epidemic exceeds the length of the data, then simulations on the static and on the temporal network have to be handled with caution as they might underestimate the size of the epidemic.

3.5 Model networks

In order to better understand the origin of the decrease of the epidemic size with augmented epidemic speed, and how it depends on properties of the temporal network, we build several model networks which change some of the properties of the temporal network.

The model networks are:

- **time shuffle** (shuffled starting time):
  Every event in the temporal network can be attributed 4 values: the IDs for the two nodes which are in contact, the starting time of the contact and its duration. For the time shuffled network, the value for the starting times of all events are randomly redistributed. Thus the same amount of new contacts are started at any time instant for the time shuffled and the original network, but the starting times for contacts between each pair of nodes have changed. By shuffling the starting times of events, we keep the contact-time distribution but lose correlation of events, like group meetings. Also the node inherent burstiness is reduced, as starting times between different events are exchanged. The

activity fluctuations are mostly preserved.

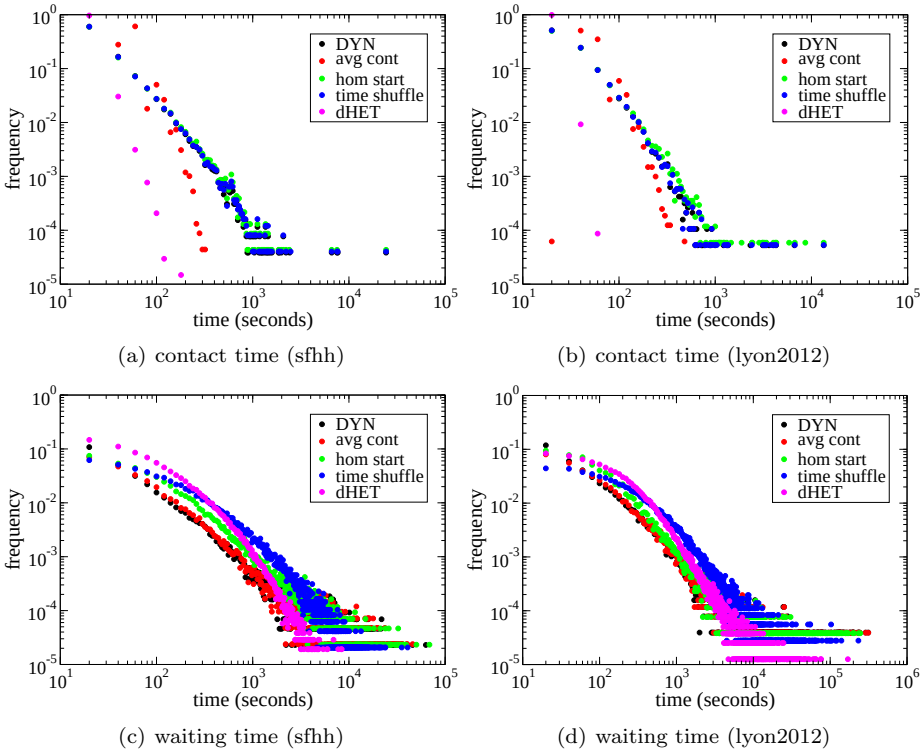
- **hom start** (homogeneous starting time):
  For the "hom start" network, the starting times of all events are modified. To this end, events are ordered according to their starting time. The order among events starting at the same time is random. Then all starting times are distributed at equal distance in time. The starting time difference between two consecutive events is equal to the length of the data set divided by the number of events. The order of events is conserved. Setting starting times of events at a homogeneous distance will mostly eliminate the activity fluctuations, while keeping the contact time distribution. It will also preserve some of the node inherent burstiness, as two events from one node, which start within a short time, will stay comparably close together, depending on the overall activity of the network between the two starting times.

- **avg cont** (average contact time):
  The contact-time distribution is changed by setting the length of all contact durations to the average contact duration $\langle t \rangle$ while keeping the exact starting times of the events. Since the average contact duration is not an integer multiple of the timestep length, for each event the length of the average contact duration was chosen at random to be either rounded up or down to the next multiple of 20 seconds (the duration of the network's time steps). The probability of rounding up to the next integer multiple of 20 seconds was chosen proportional to the difference of the average contact duration and the next lower multiple of 20 seconds divided by 20 seconds, $p = \langle t \rangle/20 - \lfloor \langle t \rangle/20 \rfloor$. This transformation also changes the weight distribution of the corresponding aggregated HET network. Weights of each link are now proportional to the number of occurrences of each link.

- **dHET**:
  The aggregated static network with heterogeneous weights (HET) is transformed into a temporal network with Poisson dynamics. At every time step, for each weighted link of the HET network the link is set as active with probability given by its weight. If discrete simulation timesteps are used, it does not make a difference if we use HET or dHET for any one simulation. However, discrete simulation timesteps are only valid for low β. Therefore we use continuous simulation steps here and HET and dHET need to be distinguished.

By changing the starting position of contact events in the temporal network, the length of the temporal network was sometimes increased. In this case, we cut off contacts which were lasting longer than the original length of the network and added them at the beginning of the network. In case two events between the same nodes overlapped, the later event was moved so that it started after the first ended, preserving the total duration of contacts of the network. To build the average over several simulations all model networks were frequently rebuilt.

All four transformations slightly influence the waiting-time distribution. We plotted the distributions for one realization of each model network in Fig. 3.9. For the model with homogeneous starting times ("hom start"), the frequency of very small waiting times, which are due to clustering of contacts in time, is decreased when starting times are drawn apart, and the frequency of slightly longer waiting times is increased instead. Long waiting times are not affected. This effect is even stronger for the time shuffled version, as the starting times of events for the same node are not merely drawn apart but randomly exchanged with starting times of other events, eliminating node-inherent burstiness and also reducing the number of long waiting times. In



(a) contact time (sfhh)

(b) contact time (lyon2012)

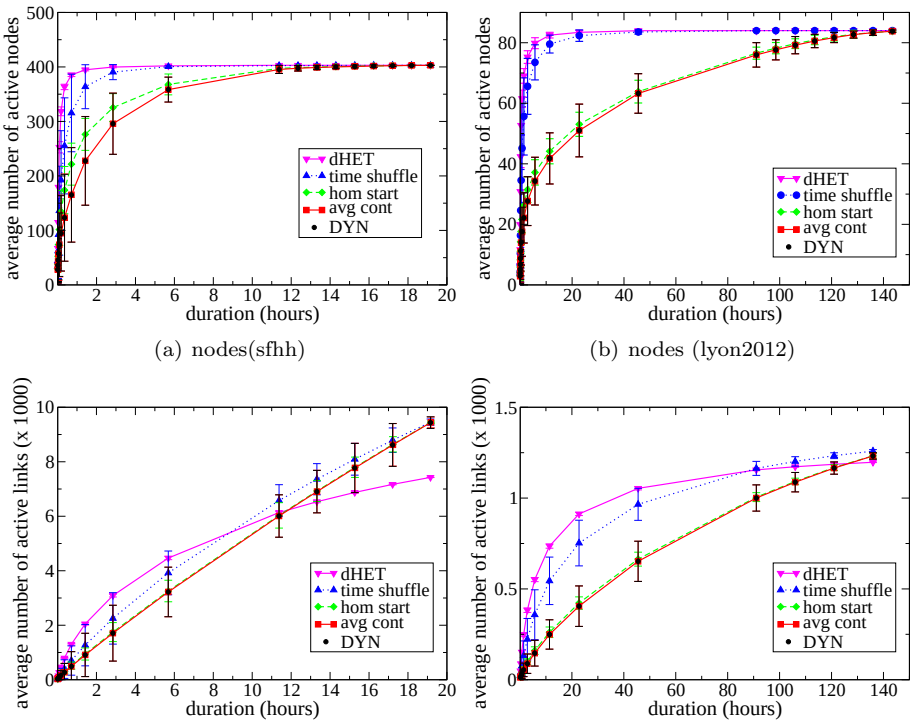
(c) waiting time (sfhh)

(d) waiting time (lyon2012)

Figure 3.9: Waiting time distribution and contact-time distribution for the temporal networks "sfhh" and "lyon2012" and the corresponding model networks

the dHET case, short waiting times are common and long waiting times are rare. Due to the Poisson dynamics, nodes are mixing more, contact times are shorter, and events are abundant and spread over the entire time of the temporal network.

The contact-time distributions are only different for the model network with average contact times and the model with Poisson dynamics. In both cases, contact times are short and not broadly distributed.



(a) nodes(sfhh) (b) nodes (lyon2012)

(c) links (sfhh) (d) links (lyon2012)

Figure 3.10: Average number of nodes and links which are active over a time period on the temporal networks "sfhh" and "lyon2012" as well as on various modifications thereof. Together with the average, the standard error is plotted.

As the limited number of present nodes and links during any time period of the network was conjectured to be responsible for the different outcomes of epidemics with different spreading velocities, here we compare these measures for the different model networks. In a temporal network, nodes arrive and leave over time, some contacts are very rare, others are quite frequent. If only a short period of the temporal network is considered, not all possible links will be active during this time period and not all nodes will have had contacts. In Fig. 3.10 the average number of present nodes and links per time window is shown. For the dynamic data, the average number of present nodes increases rapidly for short time windows, however, there is no scale at which no nodes are introduced. The average number of present nodes increases constantly.

The average number of links and nodes is identical for the temporal network and the "avg cont" network, suggesting that the correlation of the starting times of events plays a more prominent role than the actual contact length for the presence of links and nodes over time.

For the network with homogeneous starting times, the contacts are spread out more in time,

eliminating activity fluctuations. Thus at any time on the temporal network, there is a similar number of present nodes for the "hom start" network. Therefore, the standard error is lower for the "hom start" network than for networks with activity fluctuations, like the DYN network. For the "lyon2012" network, the average number of present nodes is very similar to the average number of present nodes on the original temporal network. However, for the "sfhh" network, which has very strong activity fluctuations, the average number of present nodes is slightly higher on the "hom start" network compared to the DYN network.

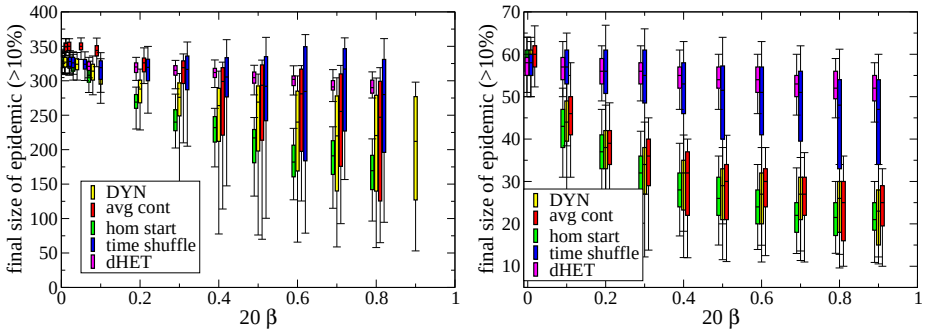
The network with shuffled events has a higher mixing of nodes. Nodes are not confined any more to a limited time period in which they are present on the network, event correlations are completely eliminated. Events including nodes can appear at any time during the temporal network. This increases largely the average number of present nodes and active links per time period.

In the case of the dHET network, the probability of appearance of a node at any time only depends on its strength and is therefore constant over the entire length of the temporal network. The average number of present nodes and links during a certain time is highest for this temporal network with Poisson dynamics. However, as link presence is stochastic, links with low probability of appearance have a high probability not to be activated in the dHET model network, as the dHET network only extends over a limited time period, identical to the length of the original data. Therefore each of the realizations of the dHET network has a smaller number of links than the other model networks.

A difference can be seen between the "lyon2012" network and the "sfhh" network. While in the "sfhh" network, most nodes have quickly made their appearance, in the "lyon2012" network the increase of the average number of nodes which have been present in a certain time period is much slower. This is related to the different settings of the dataset. While most people will stay during the entire conference rather than just coming for one or two hours, in the hospital, nurses or doctors might only come on specific days. Concerning the average number of active links, we have the opposite case. At the conference, many weak links lead to a slow increase of the average number of active links, while in the hospital, very short and rare encounters are less likely, meetings are not as random (see also Fig. 2.3(d), the distribution of weights, which is here the probability of link activation). Furthermore the high number of weak links in the "sfhh" dataset also leads to a stronger discrepancy between the total number of links for the dHET network and the aggregated temporal network.

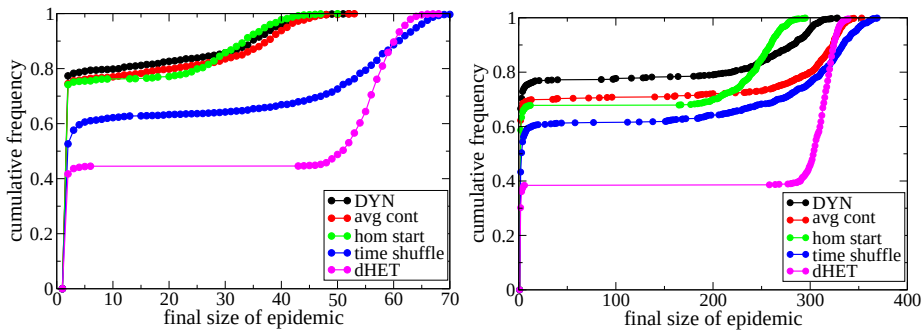
In Fig. 3.11 an SIR process was simulated for all model networks for increasing propagation parameter β, as before. For simulations on the "lyon2012" network, the final size of the epidemic decreases strongly for faster epidemics for the original temporal network (DYN), as well as the model networks "avg cont" and "hom start", which have the same average number of present nodes and a very similar waiting-time distribution. The model networks that have a high average number of present nodes even over very small durations do not show a strong difference for slow and fast epidemics. For the time shuffled network, there is only a small decrease with β and for the dHET network, the decrease with β is almost not visible at all. For the static HET network, there would be no decrease with β, as could be seen in Fig. 3.5 above.

Simulations on the "sfhh" data set look slightly different. This is mainly due to the fact that activity fluctuations are very strong. The distribution of the final cases of the epidemic is much broader for the networks with daily patterns: the DYN, the "avg cont" and the "time shuffled" networks. For the model with homogeneous starting times, the decrease with β is the strongest even though the waiting-time distribution and the average number of nodes suggest otherwise. Unlike in the temporal networks with activity fluctuations, in the "hom start" model there are no periods with very high activity and high mixing of nodes. In networks with high activity fluctuations, the epidemic either dies out quickly, or it survives up to a peak of activity then



(a) sfhh (b) lyon2012

Figure 3.11: Number of final cases for the temporal network of "sfhh" and "lyon2012" and the modifications thereof. Both network do not include nights. The boxes show the 25 and 75 percentiles as well as the median of the distribution of final cases where only epidemics which reach more than 10% of the network are considered. The whiskers mark the 5 and 95 percentiles.



(a) lyon2012 (b) sfhh

Figure 3.12: Cumulative distribution of final cases for epidemics with $\beta = 0.015$. (a) For the "lyon2012" data set: $\mu = \beta/100$. (b) For the " sfhh" dataset: $\mu = \beta/50$. Both datasets do not include nights.

spreading widely and quickly among all present nodes. Thus the epidemic outcome can often have a high number of final cases as well as a low number of final cases. For the model with average contact times, the final size of the epidemic is much higher than on the temporal network. Possibly, because the probability of transmission on a contact event is a concave function of contact time. The transmission probability over two equally sized contact events is thus higher than over one of very short and one of very long length. Also, if the "avg cont" network were aggregated, the distribution of link weights would not be the same as on the dynamic network.

In the cumulative distribution of final cases for $\beta = 0.015$ (see Fig. 3.12) the difference between the model networks and also between the two data sets become easily visible. Simulations are started at a random time, independent of the presence of the seed in the network at that time. The extinction probability is highest for the original temporal network due to the burstiness of the dynamics linked to the high activity fluctuations. The epidemic can often end even before the infected seed has had its first contact in the network. Simulations on the network with homogeneous starting time, on the other hand, end less often before infecting other nodes, as activity is homogeneous and there are always nodes present which can be infected. For the network with Poisson dynamics, epidemics have the lowest extinction probability. The distribution of the final number of cases is very broad for networks with strong activity fluctuations, like DYN, "time shuffle" and "avg cont", and therefore the cumulative distribution increases gradually for higher case counts of the epidemic, while for the more homogeneous networks without activity fluctuations, like "hom start" and "dHET", a characteristic number of final cases can be observed as a sharp increase in the cumulative distribution function. For the "sfhh" dataset, this increase is at a medium final size of about 250 final cases for the network with homogeneous starting times. Outcomes in this range or below are therefore highest for the "hom start" case. For the dHET network, this peak lies around 315 final cases. In networks with high activity fluctuations, smaller outcomes are most likely, but large outcomes can be reached as well.

For the "lyon2012" network data, these tendencies are not as strong, as activity fluctuations are much lower. However, the shuffling of starting times has a much greater effect, as nodes do not mix as widely as in the conference data.

All in all, the main effect for the decrease of the epidemic size for faster epidemics is due to the reduced number of nodes and links that are active at any given time span. Networks with a high turnover of nodes and links are more susceptible to this effect. Activity fluctuations, on the other hand, are not responsible for the decrease of the epidemic size with the epidemic speed, even though they can change the shape of the distribution of final cases.

3.6 Conclusion

We have seen that patterns at small time scales have a stronger influence on epidemics with high probabilities β and μ than on those with low β and μ. The size of the epidemic is reduced in those cases of fast epidemics mainly because not all nodes can be reached on the network, but also because the number of links between nodes is limited to those created through interactions taking place during the epidemic. The introduction of nights and long times without activity can be another reason for the reduction of the final size of the epidemic. However, in spite of all the variations of the epidemic size due to temporal properties, static networks can be a good first approximation even of bursty temporal networks, if they are aggregated over the length of the network on which the epidemic takes place. In the next chapter we will look more closely under which conditions the resolution of the temporal data can be reduced, extending this simplification also to the structural resolution of static networks.

Chapter 4

Data representation

With the advent of new technologies, ever more detailed data sets can be obtained. However, for large scale simulations, data sets with very high precision do not exist yet and also would quickly become unmanageable. Therefore, large-scale epidemic models are usually informed by general data without high spatial or temporal resolution. This can be census or demographic data, data about co-location, shared office spaces, human mobility or the like [97, 21, 56, 5, 17]. Furthermore, a high level of detail can easily obscure the important aspects of the data. Especially, if results are expected to be generalizable, it can be helpful to filter out unnecessary details and only keep those aspects of the data which have a major influence on the outcome of the epidemic. Aspects of the networks, like the degree distribution or cluster coefficient, which are known to have an influence on epidemic spreading should be conserved when the data is simplified.

As the high-resolution, temporal contact data contain many aleas which are situation specific, they are difficult to generalize. In order to be valid for simulations with a different amount of participants, at a different time or place, a representation of the data needs to be found which is independent of these specific details and only retains those properties of the data which are general for different situations.

For simulations on a large scale, data might not need to be as precise. The level of precision needed on the data is related to the level of precision which is wanted for the outcome of the epidemic and also to the sensitivity of the process on the data to the details of the data. For example, to evaluate targeted vaccination strategies for different groups, information on the risk of infection for each individual is unnecessary. Also, slow spreading processes seem to be less sensitive to data fluctuations on small time scales. It is therefore plausible that for fast processes data need to have a higher temporal resolution than for slow processes.

When dealing with high resolution data sets, it is therefore an important task to find a way to modulate between data representations of different levels of detail and to find the right amount of information that needs to be retained in the data [14] given a specific problem or task.

We will construct here different representations of the data with varying levels of detail. The data will be more and more simplified by replacing exact data by distributions or averages. In Sec. 4.2 we look for the optimal resolution in time depending on the process that is run on the data. To this end, we compare the outcome of epidemic simulations on temporal networks whose time resolution is reduced through aggregation. By aggregating temporal events into more general weight information on the links, the exact timing of events is replaced by a Poisson distribution, which disregards temporal order but retains the contact probability of links. Other distributions, formed by processes with memory and keeping information on the burstiness of the contacts and the contact-time distribution, could be considered at a later time to obtain better

approximations.

Once the temporal resolution is decided upon, for each time the corresponding static network graph can be simplified as well, depending on the level of detail that is considered necessary for the epidemic simulation. We consider simplifications on two levels, the distribution of weights and the placement of weights on the network. Nodes can completely lose their identity or be assigned to different groups and only retain a shared group identity, where the group identity comprises the aggregated properties of the contained nodes, like their average degree, average strength or average clustering coefficient. We discuss the importance of a proper choice of groups in Sec. 4.2.1. Together with the suppression of individual properties of the nodes, information on the exact weights of the links between groups can be replaced either by providing each link with the average weight or by drawing weights from a weight distribution. The former leads to a very basic network that can be represented as a contact matrix. The contact matrix representation of data is widely used in epidemiology [72, 93, 109, 40, 42, 66, 89, 43, 29] as it constitutes an improvement over the homogeneous mixing hypothesis and can easily integrate data which lacks higher levels of precision. Being easily generalizable and adaptable to different situations, it furthermore lends itself to predictions for situations for which actual data cannot be obtained. However, as we will see in Sec. 4.2.2, the heterogeneity of weights can be significant for the epidemic simulations and discarding it may lead to wrong predictions. We therefore introduce the contact matrix of distributions, which replaces the exact weights by weights drawn from a negative binomial distribution. Negative binomial distributions can account for the broad distribution of weights and also the high number of zero-weight links. They are often used to model contact distributions [58, 42, 66]. By simulating epidemics on the various simplified data representations, we can compare the influence of the respective modifications of the networks constructed from the data representation and assess the level of detail needed for the purpose of obtaining correct predictions of the probability for a member of one of the groups to become infected, the influence of different groups on the spreading (see Sec. 4.2.4) and the overall size of the epidemic.

The importance of roles in the network is influenced by their characteristic weight distributions. Depending on how well these distributions can be approximated by the average weight, the relative importance of the groups changes for the contact matrix representation. Approximating the weight distribution by the average weight increases on average the transmission probability. We account for this by adapting the average number of secondary infections per node in Sec. 4.2.5. Nevertheless, adapting R_0 cannot account for the change of group importance in the contact matrix representation as compared to the dynamic network. We can therefore conclude that a global rescaling of the model parameters is not able to compensate for heterogeneity induced differences between the groups which have a direct influence on the outcome of the epidemic. Depending on the properties of the data and the distribution of weights among the groups, it can therefore be crucial to include information on the heterogeneity of weights in the contact matrix representation. Even in cases where the contact matrix is an acceptable approximation of the exact network, the contact matrix with distributions will help to obtain simulation results with higher accuracy.

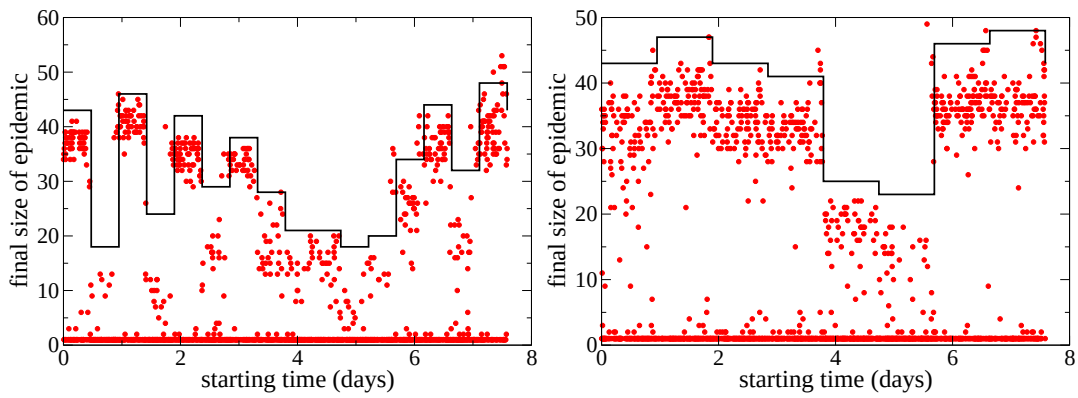
4.1 Time resolution

High-resolution temporal networks over a long time period can become quite large. However, not all processes on the network might need the same level of precision of the data. As seen before (Ch. 3), fast spreading processes sense more details of the data than slow processes. It is therefore likely that slow processes do not require the same level of detail of the data as fast

processes. A resolution of 20s per timestep might therefore be an overkill, and data with lower resolution could suffice for slow processes. Nevertheless, collecting data with a low resolution can also pose problems, even for slow processes, if the exact contact length is neglected, as seen in Sec. 2.6. Measurement needs to be precise concerning the duration links are active, whereas the exact timing of events plays a secondary role. Integrating the exact contact time length into data with lower time resolution using weights can be a good compromise between high resolution data and data with lower resolution. However, depending on the spreading parameters, a complete loss of temporal information can lead to an overestimation of the outcome of the epidemic (see Sec. 3.4), even if weights are included. When the data is bursty, some temporal information seems necessary and aggregating over the entire network, may lead to biased results. Nevertheless, simulations on static networks can be a feasible alternative to simulations on the time-resolved data if the duration of the epidemic coincides with the aggregation time of the static network. For specific parameter sets it has been furthermore shown that aggregating over daily networks is sufficient to obtain similar results as with the full temporal resolution of the contact network [88]. It is therefore conceivable that a minimum temporal resolution exists so that increasing the resolution further will not add information which influences the epidemic process on the network. We are looking here for such a minimum timestep size, depending on the speed of the epidemic.

In order to find such an optimal aggregation time step, we simulate epidemics using the SIR model on the "lyon2012" dataset. Nights were not removed. The data are reduced to a length which is a power of 2 and then aggregated with different time step sizes, using heterogeneous rescaling with weights as described in Sec. 2.6.

By increasing the time-step length over which the temporal network is aggregated, we can modulate between the high-resolution temporal network and the heterogeneous static network (HET). The temporal fluctuations of the data decrease with larger aggregation time steps. Whereas networks with aggregation times below 12 hours can still show daily patterns, aggregating over entire days will make daily patterns disappear.



(a) aggregation over 11.4h

(b) aggregation over 22.8h

Figure 4.1: The final size of the epidemic as a function of the starting time for two networks with different aggregation step length. The black lines signal the number of active nodes in each timestep. Simulation is done for $\beta = 0.04s^{-1}$ and $\mu = 0.0004s^{-1}$

In Fig. 4.1 the influence of the starting time on the outcome of the epidemic is shown for very fast epidemics on two networks with different temporal resolution. The daily patterns, which are still present in the network with a resolution of 12 hours, strongly influence the outcome of the very fast epidemic. The network which is aggregated over 23 hours does not show any daily

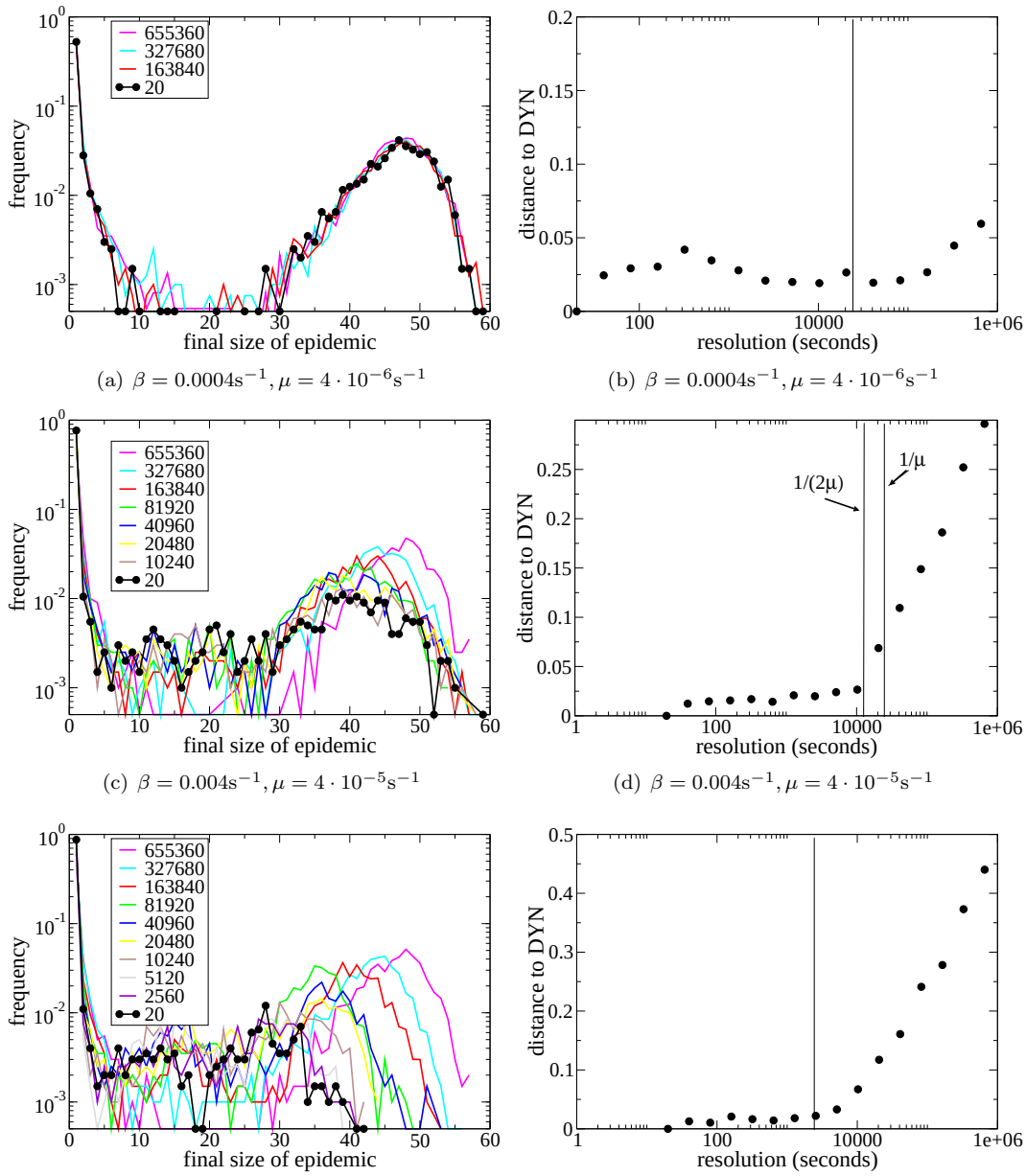
patterns anymore, however, weekly patterns, in the form of a strong reduction of activity during the weekend, are still visible.

We test epidemic spread for three parameter sets corresponding to a very fast, medium and slow epidemic on temporal networks of different time resolution. The distribution of final sizes of the epidemic is plotted in Fig. 4.2. Up to a certain aggregation step length, the distribution of case counts does not change. Only when the aggregation step length is of the order $1/\mu$ does the average number of people reached by the epidemic increase. Nodes are on average infectious for a duration of the order $1/\mu$. In this time they can infect any of their neighbors. When the network is aggregated, the average number of neighbors each node is in contact with increases (see Fig. 2.5(a)). If the aggregation time is larger than the infectious time, then the node can infect more neighbors in the aggregated than in the temporal network. In Fig. 4.2(b), 4.2(d) and 4.2(f) the L^2-norm between the histogram for the spreading on the original data and the histogram for the spreading on the aggregated data is shown. The L^2-norm is used as an approximation of the distance between the two distributions. In the appendix (Fig. A.1) we have used Pearson's χ^2-test to decide for which aggregation timesteps the distributions are significantly different. In a first approximation, the distributions become very different for aggregation steps greater than $1/\mu$. Possibly the variability of the network at different timescales also plays a role. If, for example, the neighbors for a specific node do not change much over time, and if with longer aggregation no new information is added to the network, then aggregating over timesteps longer than $1/\mu$ will not have a great impact on the outcome of the epidemic. If on the other hand there is much difference between two subsequent instants of the aggregated temporal network, it can be important that the network is aggregated over the exact time of infection of a node. If the node gets infected close to the end of an aggregation interval, it effectively can infect nodes which are neighbors in both intervals, the interval in which it got infected and the subsequent one. Therefore an aggregation timestep smaller than $1/\mu$ can be necessary. This can be seen for the simulation with $\beta = 0.004$ and $\mu = 0.00004$. The aggregation timesteps corresponding to $1/\mu$, which are around 7 hours long, divide the temporal network into static snapshots with very different numbers of nodes due to the strong night-day patterns of the data. Here, and also for slightly longer aggregation timesteps (compare Fig. 4.1(a)), the daily patterns can still have a big impact, greatly changing the graphs from one instant to the next. In this case smaller aggregation timesteps might be necessary. Furthermore, for a higher probability of propagation, being infectious in more than one time instant of the temporal network has a higher impact than for a lower propagation probability. Therefore, aggregation timesteps below $1/\mu$ can be necessary in those cases as well.

We finally note that it could make sense to not aggregate the temporal network in equal timesteps but to use smaller timesteps in regions of high variability of the data and longer timesteps during night periods or periods when the data does not change much.

4.2 Structural resolution

Every instant of the temporal network can be described as a static network. By aggregating over longer time periods, the instantaneous static networks become more complex, gaining nodes and links with weights. In addition to a simplification in time, we can also simplify the structure of the static networks. In order to reduce the information content for the static network, exact data must be replaced by average values or distributions. This reduction of information can be compared to lossy data compression. The aim is to compress only those parts of the data which are irrelevant for the task at hand and to keep the information in the data which is essential for epidemic spreading.



(a) $\beta = 0.0004\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-6}\mathrm{s}^{-1}$

(b) $\beta = 0.0004\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-6}\mathrm{s}^{-1}$

(c) $\beta = 0.004\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-5}\mathrm{s}^{-1}$

(d) $\beta = 0.004\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-5}\mathrm{s}^{-1}$

(e) $\beta = 0.04\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-4}\mathrm{s}^{-1}$

(f) $\beta = 0.04\mathrm{s}^{-1}, \mu = 4 \cdot 10^{-4}\mathrm{s}^{-1}$

Figure 4.2: Left column: distributions of the final size of the epidemic for networks with different aggregation time steps for three sets of parameters with $\beta/\mu = 100$. Right column: L^2-norm between the distribution of the final size of the epidemic for the network with highest temporal resolution and networks with lower temporal resolution as a function of the aggregation time step. The vertical lines mark $1/\mu$.

	average weight	weight distribution or exact weights
all weights	FULL	HETshuf
all weights (groupwise)	CM	CMD
nonzero weights	HOM	HETwshuf

Table 4.1: Data representations categorized according to the information on weights. The rows specify if all weights, all weights within and among groups or only the non-zero weights are concerned, keeping the link-structure of the network. The columns specify if these weights are replaced by their average, by randomly drawn weights from the weight distribution (CMD) or if weights are shuffled (HETshuf, HETwshuf).

	set1	set2
β	60/day	240/day
γ	1/day	2/day
μ	0.5/day	1/day

Table 4.2: Parameter sets which are frequently used in the following.

We are interested in two kinds of information on the static network: the weights and the placement of weights on the network.

In a first approximation, the network structure is defined by the presence and absence of links. In a second approximation, it can make sense to consider the exact placement of the weights as part of the network structure. Often, the placement of links on the network comes about by applying a threshold to a fully connected network with different weights. Links with weights below a certain threshold will be removed. The information of the placement of weights on the network, including zero-weighted links, therefore integrates all the topologically possible networks with different weight thresholds.

We can simulate the effect of losing all information of the weight placement and keeping all information of the exact weights, by shuffling the links of the network. We test for two possibilities, shuffling all weights, including zero-weighted links or only shuffling the non-zero weighted links. Concerning the information on the exact weights, we consider two different levels of information loss. Replacing all weights with the average weight corresponds to a strong loss of information, using a distribution which approximates the original distribution of weights corresponds to a weak loss of information.

The networks which are created by these data reduction methods are:

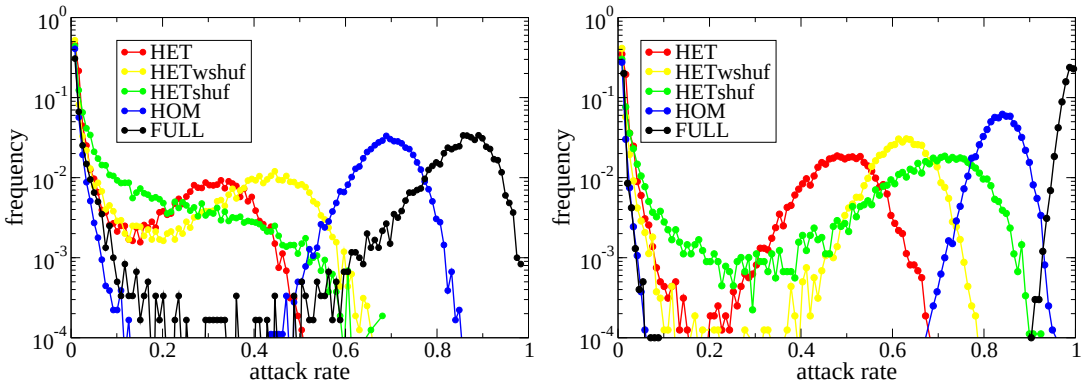
- HET
  The heterogeneous static network, as described before, which integrates all temporal information in the form of weights. It is the static network with the most available information, so that we use it as comparison for the other static networks.

- HETwshuf (weight-shuffled HET network)
  All non-zero weights of the HET network are shuffled. The link structure of the network is identical to the HET network and all information on the weight distribution is kept. However, information on the exact placement of the weights is lost.

- HETshuf (shuffled HET network)
  All links of the HET network are shuffled to random positions. Thus, all the information on the weight distribution is kept, and all information on the placement of weights and links is lost.

- HOM (homogeneous static network)
  The network structure is identical to the HET network but the weights of all non-zero weighted links are replaced by the average of the weights of the non-zero weighted links in the HET network.

- FULL (fully connected network)
  The weights of all links are identical to the average weight of all links in the HET network, including the zero weight links. This network loses thus all but the average information on its weight distribution and keeps no information on its structure. The full network corresponds to the homogeneous mixing assumption, which is often used as a first crude approximation for contact patterns among people.



(a) set1 (b) set2

Figure 4.3: Distributions of the final size of the epidemic for static networks with varying levels of detail in the underlying data as described in the text. Parameter sets see Tab. 4.2.

To compare the effects of the different simplifications of the data, we model an SEIR epidemic on the networks, using two parameter sets: set 1 for a slightly slower and set 2 for a slightly faster epidemic (see Tab. 4.2). In Fig. 4.3 the distribution of the final sizes of the epidemic is shown for the simulations on the networks created from various representations of the "obg" dataset. The effect of the loss of precise information on the data is clearly visible.

Losing all information, both the information on the distribution of weights and the information on the network structure, as is the case for the FULL network, results in the least similar outcome of the final sizes of the epidemic compared to the HET network. Keeping, on the other hand, both the information on the weight distribution and on the network structure and only losing the information of the placement of weights on the links (HETwshuf) has the most similar outcome to HET of the compared models.

While complete loss of information on the placement of the links and complete loss of information on the weight distribution both have a severe influence on the outcome of the epidemic, the weight distribution seems to be slightly more important in this case. This can be seen by comparing the distribution of final sizes of the epidemic of the HET network with either the distribution of final sizes of HOM, where information on the heterogeneity of weights is lost, or with the distribution of final sizes of HETshuf, where information on the placement of links is lost. The distribution of HETshuf is closer to the distribution of HET than the distribution of HOM. However, both variants constitute a simplification of the original data which is too severe.

Predictions based on these data representations would be too different from predictions based on the HET network in order to represent valuable replacements for the latter.

4.2.1 Choice of groups

In order to keep some information on the structure of the placement of weights on the network, instead of replacing the weights by distributions or by their average on the entire network, the same procedure can be applied only on parts of the network. To this end, we assign the nodes of the network to different groups or classes. Within each group, nodes lose their individuality. Just as with the entire network before, links are placed randomly between and among nodes of the different groups. However, the average weight or the weight distributions are now calculated separately for each group and for links between groups. This way, the structure of high density and low density clusters in the network can be partly preserved.

By changing the number of classes into which nodes are arranged, we can tune the network between the HET network, which contains the complete available information of a static network, and either the FULL network (in the case of average weights) or the HETshuf network (with exact, but shuffled weights), or rather a network which resembles the HETshuf network but with a fitted weight distribution. If every node is its own class, all information is kept, whereas all structural information is lost when all nodes are grouped into the same class.

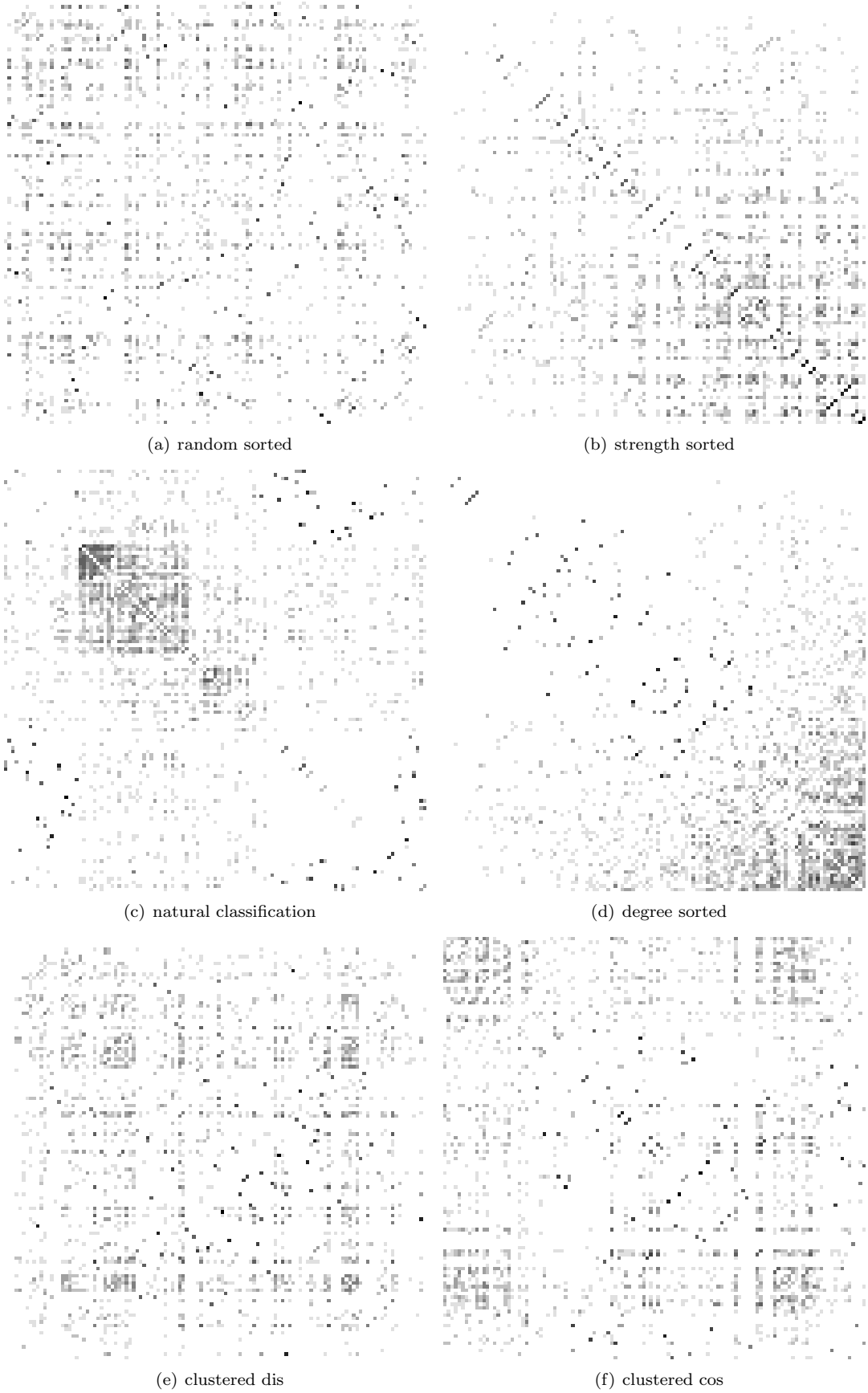
When nodes are assigned to different classes, they lose their individual properties like degree, strength or choice of neighbors. Links within and among groups are placed randomly. Individual properties of nodes are replaced by group properties. The information content of the network will therefore not only depend on the number of groups but also on the choice of groups. The weight on each link is replaced by another value, either the average or a value drawn from a distribution. The closer the new value is to the old one, the better is the approximation. If weights within a group or between two groups are already fairly homogeneous, then replacing them by random values drawn from the corresponding distribution will not severely alter neither the information on the weights nor on the structure of the network. An intelligent choice of groups could therefore be a choice in which link weights among and between groups are fairly homogeneous. Furthermore the number of groups should be neither too large, so that a reasonable reduction of detail is obtained, nor too small, so that the outcome of epidemic simulations will not deviate too much from the outcome on the HET network. If weights of links among and between groups are homogeneous, then the adjacency matrix with weights should be close to a block form.

We show in Fig 4.4 the adjacency matrices obtained for the following methods used to group nodes together:

- **random**
  Node placement on the adjacency matrix is random (Fig. 4.4(a)). If nodes are assigned randomly to groups, the adjacency matrix does not show any particular structure.

- **degree**
  Nodes were sorted according to their degree. In the adjacency matrix, they are placed next to each other with increasing degree from left to right and top to bottom (Fig. 4.4(d)). Nodes with similar degree could then be grouped together.

- **strength**
  Nodes were sorted according to their strength. In Fig. 4.4(b) they are placed next to each other with increasing strength from left to right and top to bottom. Nodes with similar strength could be placed in the same group.



(a) random sorted

(b) strength sorted

(c) natural classification

(d) degree sorted

(e) clustered dis

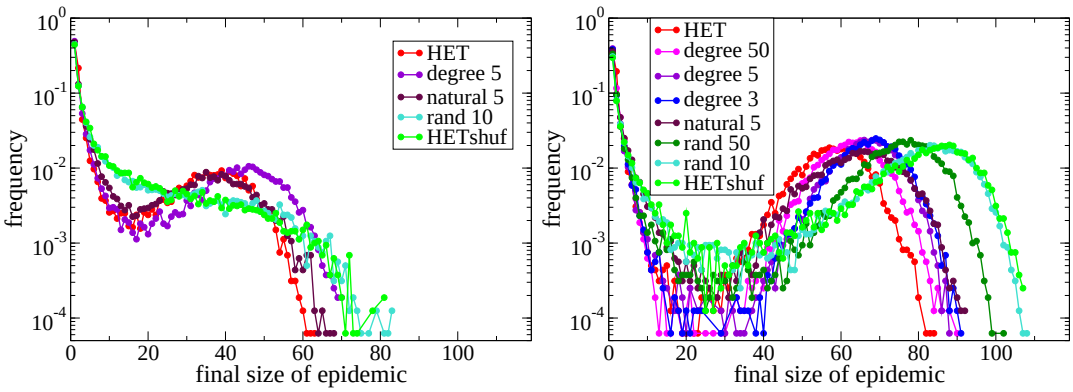
(f) clustered cos

Figure 4.4: Adjacency matrix of the nodes, ordered so that nodes which are next to each other can be classed in the same group. The grey value is proportional to Log(1+weight).

- **clustered**
  An algorithm to cluster nodes with similar link properties was tried in Fig. 4.4(e) and 4.4(f). To this end, each node was represented by a vector of its weights with all other nodes. Then, an auxiliary network was formed of these vectors in which nodes were connected by links weighted according to the similarity. For Fig. 4.4(e) this similarity was calculated as the inverse of the L^2-norm between the vectors of the two nodes (clustered dis). For Fig. 4.4(f) the cosine similarity was used instead (clustered cos). In this auxiliary network, nodes which have similar connections with third party nodes are linked by high weights. Finding clusters of nodes on this network ideally would result in groups with similar connection properties. To find the clusters, a simple clustering algorithm was run and a dendrogram was built. Nodes in the matrix are plotted in the order of the dendrogram's leaves. Thus, nodes with similar neighborhoods should end up close together.

- **natural classification**
  Since the data comes with meta information concerning the nodes, the natural classification (Fig. 4.4(c)) is the one which groups nodes of the same role class together, as given by the meta information on the data. Nodes, placed from left to right (and top to bottom) on the adjacency matrix, belong to the following groups: Caregivers, Assistants, Nurses, Doctors, and Patients. The sizes of these role classes are given in Tab. 4.3.

The block structure appears most prominently for the natural classification and can be imagined with much good will for the degree sorted nodes. Also, for all practical purposes, it is favorable if groups are chosen by local properties which are independent of the network structure as a whole in order to be more stable in case of minor network changes.



(a) set 1 (b) set 2

Figure 4.5: Histograms of the final size of the epidemic for networks with groups of different sizes and formations from 16000 simulations. Every 1000 simulations random groups were newly built, every 100 simulations links between and among groups were shuffled.

In order to assess the influence of the group selection mechanism on the structure of the network representation and the corresponding outcome of simulations performed on this network representation, we create different network representations in which nodes are either randomly assigned to different groups or grouped together by degree. We also consider a partition of the network into groups according to the roles participants fill in the hospital. The exact weights are preserved. This way we can test for the influence of group number and grouping method only. All links, including those with zero weight, between two groups as well as all links within

Assistants	10
Doctors	20
Nurses	21
Patients	37
Caregivers	31

Table 4.3: Group sizes for the roles in the hospital data set obg

one group are shuffled randomly. Group sizes for the randomly assembled nodes were chosen homogeneously. For the network with degree sorted groups, nodes were sorted according to their degree, then a sequence (n1,n2,n3..) of group sizes was chosen, and the nodes with highest degree were placed in group n1, nodes with subsequently lower degree in group n2 and so on until all nodes are placed into groups. Since the degree distribution has few nodes with very high degree and many nodes with very low degree, a partition into homogeneous groups seemed less able to optimally group nodes with similar degree together. An optimal partition could be conceived as one in which the new degree distribution, after reshuffling of links in each group, approximates the original degree distribution best for a given number of groups. As nodes in one group after reshuffling of links have a degree close to the average degree of the group, a possible partition into groups can be achieved by optimizing the approximation of the area under the function which plots the degree of nodes depending on their position in the ordered list with a given number of rectangles,where the base corresponds to the size of the group and the heights to the corresponding average degree in the group. However, due to a lack of time this algorithm was not tried out. Instead, group sizes were chosen in order to make the result comparable to the case where groups are chosen according to the roles in the hospital. For the case with 5 groups, the same group sizes were chosen for the natural groups and the network with degree-sorted groups. For the natural groups, the group sizes are shown in Tab. 4.3. The group sizes for degree sorted groups with three or ten groups were chosen slightly arbitrarily but smaller group sizes were chosen for high degree nodes and larger group sizes for low degree nodes. When the network was partitioned into three degree-sorted groups, the group sizes were: 20, 31, 68. For the network with ten degree-sorted groups, group sizes were: 4, 6, 8, 10, 10, 13, 15, 16, 17, 20.

On these networks an epidemic is simulated using the SEIR model. Fig. 4.5 shows the distributions of the final size of the epidemic for simulations on the various networks. The HET network integrates the complete static information and is therefore taken as a reference here. When nodes are grouped together by similar degree, the outcome of the epidemic is much closer to the outcome on the HET network than when nodes are grouped together randomly. Any individual information of nodes within one group concerning the link placement is lost. The degree distribution of nodes within each group is effectively replaced by a Poisson distribution with the same average. Even when on average only two nodes are grouped together randomly ($\sim$ 50 groups), enough information on the network structure is lost in order to greatly alter the distribution of final sizes of the epidemic. Grouping random nodes together and randomly exchanging all their links with all other nodes in the graph efficiently randomizes the network, and clustering will be quickly dissolved. Grouping nodes of similar degree together will partly retain the degree distribution in the network. The outcome for the natural groups is closer to the HET network than the outcome for the degree sorted groups of the same size. Thus other properties than the degree, which are retained in the choice of natural groups, play an important role as well.

Even though the number of groups matters, intelligently grouping nodes together plays a much bigger role than changing the resolution by adding more groups. However, it is possible that for much larger networks more groups are necessary to obtain results similar to those obtained on

the corresponding HET network. The efficient number of groups and group sizes as well as the ideal criterion to select nodes for each group depending on the network structure is an issue that still needs further investigation. Possibly, community detection algorithms will be a good way to select different groups.

Nevertheless, the roles people occupy seem to represent very good markers to distinguish different groups, as every role in a hospital comes with a certain set of tasks, which imposes distinct behavioural patterns on its members. In this sense, belonging to a certain role is a single property, independent of the network structure. It comprises many different features, which otherwise might need to be evaluated separately and with considerably more effort, in order to select people for different groups. We will therefore continue only using the natural groups.

4.2.2 Heterogeneity of weights

Once a set of groups has been assigned to the data, we can apply the simplification methods described above, either replacing the weights within and among groups with the corresponding average weight or with weights drawn from a weight distribution. Network representations which are built this way are:

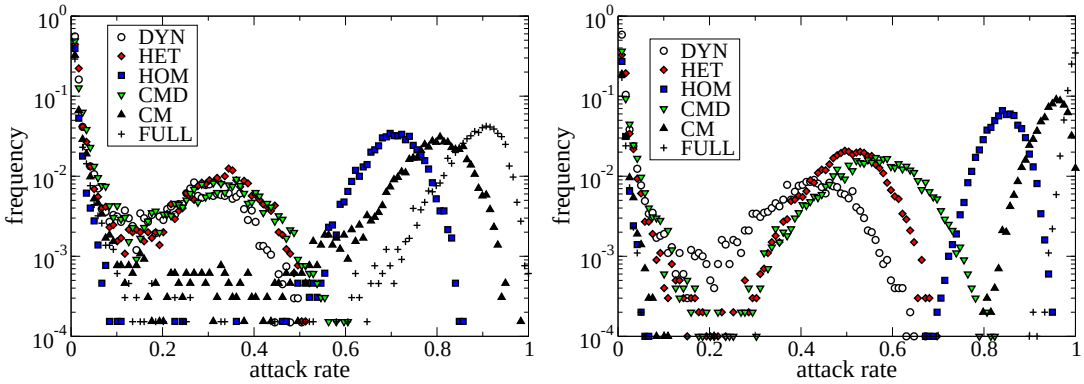
- CM (contact matrix)
  Weights, including zero weights, within groups and among groups are replaced by their average. For each group, no information on the link structure or the weight distribution is kept. However, groups have different average weights and thus at a global level some information on the weight distribution and the weight placement is kept. The contact matrix is traditionally used in epidemiology as an improvement over the homogeneous mixing assumption.

- CM0 (contact matrix with zero weights)
  Weights within and among groups are replaced by the average weight of the non-zero weighted links. The correct number, but not the placement, of zero-weights within and among each group is preserved. This data representation keeps information of the link density and average degree on a group level.

- CMD (contact matrix of distributions)
  Weights within and among groups are replaced by weights drawn from negative binomial distributions. The distributions are fitted to the original weights (including zero-weights) using maximum likelihood estimation. The fit to the weight distribution is plotted in the appendix (Fig. A.2). Thus, this data representation keeps most information on the weight distribution on a local and global scale. However, information on the placement of the weights is only retained at the group level.

The contact matrix is shown in Tab. A.1 and the corresponding parameters for the fit of the distributions are shown in Tab. A.2 and A.3. The contact matrix is traditionally used in epidemiology as a first improvement on the homogeneous mixing assumption of nodes (FULL). The contact matrix of distributions allows us to take the broad fluctuations of contact durations in the data into account.

In order to test for the influence of the weight distribution, we simulate an SEIR epidemic on the above mentioned networks. The distribution of the final size of the epidemic in Fig. 4.6 shows the strong influence of the heterogeneity of weights. While the data representations which keep the heterogeneity of weights (HET,CMD) are very similar and also approximate the DYN network best, the networks without information on the heterogeneous weight distribution (CM,



(a) set1 (b) set2

Figure 4.6: Distributions of the attack rate of the epidemic for different data representations. The "obg" data set was reduced to the first 7 days. Simulations are done for the two parameter sets (see Tab. 4.2).

HOM,FULL) all overestimate the outcome of the epidemic. The overestimation for the outcome of the epidemic when using the HOM representation is only due to the loss of the heterogeneity of the weight distribution. The exact placement of links is retained. In contrast to the HOM representation, the CM representation has slightly more information on the weight distribution, as weights are replaced by the average weights of links between or among groups. However, the full heterogeneity of the weight distribution is not retained, nor is the number or placement of zero-weighted links. All information on the exact placement of links and the average degree of each group is lost. Especially the very few but very long contacts between the Patient and Caregiver groups lead to groups with high strength but low degree. As the CM representation drops the information on the degree, spreading in those groups is much more efficient than in the original data. Thus, using the CM representation overestimates the epidemic outcome more severely than using the HOM representation.

If however the network is reduced to only contain 4 groups, by discarding the group of caregivers and all interactions with this group, then the CM representation leads to results comparable to the CMD and HET representation of the data (see Fig. 4.7). The distribution of weights between patients and caregivers, which was characterized by very few links with very high weight, disappears with the caregiver group. For the remaining groups, the average strength as used in the CM representation is a better proxy for the average degree. However, CMD still performs better than CM.

The timing of the epidemic (Fig. 4.8) for the different networks is similar if only those runs are considered which reach more than 10% of the population. As has been observed before for a non-structured population [88], the timing seems to be a fairly robust characteristic.

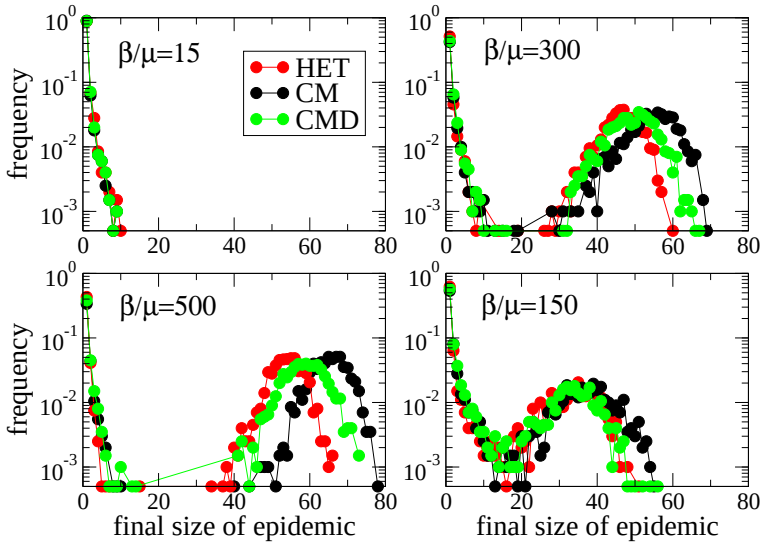
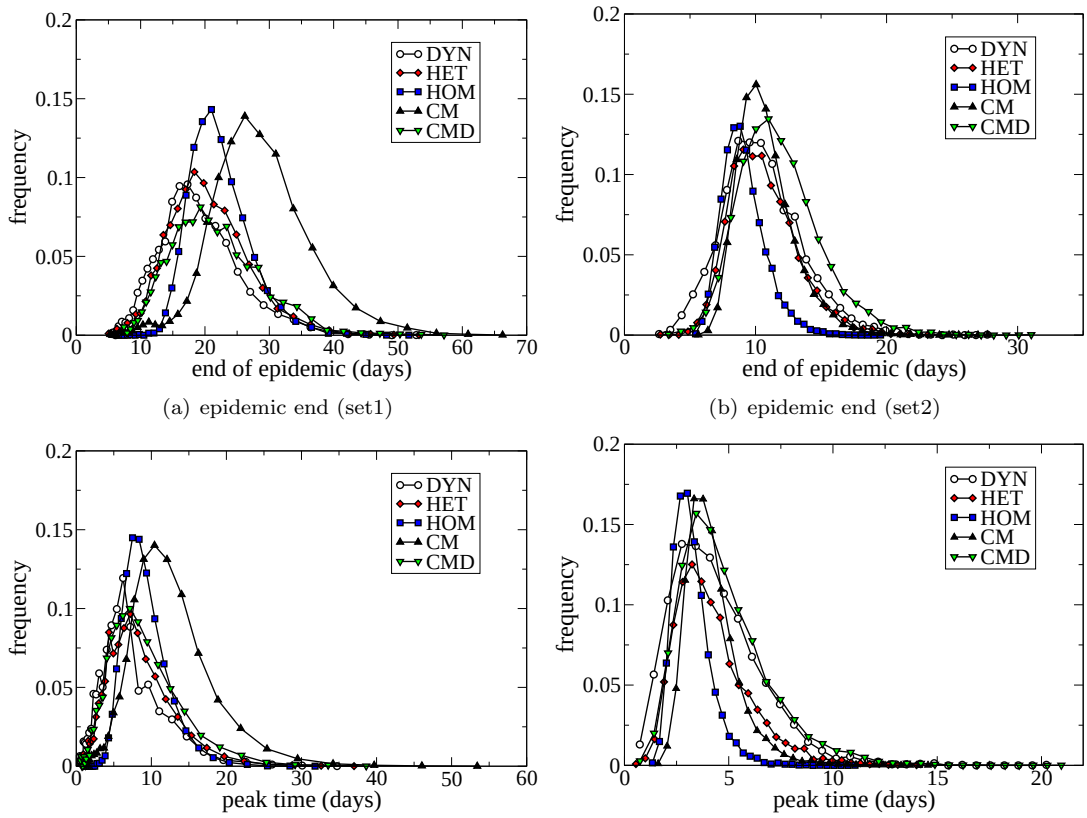


Figure 4.7: Distributions of the final size of the epidemic for the HET,CM and CMD network, based only on a network including only assistants, doctors, nurses and patients, for 4 sets of parameters: $b = 0.0694s^{-1}, r = 1./2days, m = 1./1day$, (a) $\beta = b/4, \gamma = 20r, \mu = 5m$, (b) $\beta = b, \gamma = r, \mu = m$, (c) $\beta = (5/3)b, \gamma = r, \mu = m$, (d)$\beta = (5/2)b, \gamma = 20r, \mu = 5m$.



(a) epidemic end (set1)

(b) epidemic end (set2)

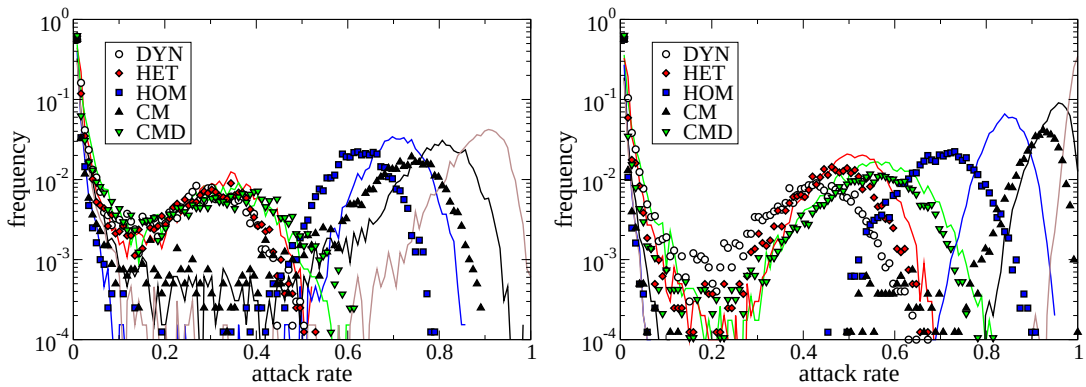
(c) epidemic peak (set1)

(d) epidemic peak (set2)

Figure 4.8: Histogram of the epidemic duration and peak time for epidemics that reach more than 10% of the population. The parameter sets are described in Tab. 4.2.

4.2.3 Daily networks

The static networks above had been aggregated over the complete dataset. Especially for parameter set 2 this approximation is too coarse. As seen before, the spreading parameters suggest rather an aggregation timestep of more or less one day. We therefore aggregate the dynamic graph on a daily basis. For each day, the network now contains a static graph with heterogeneous weights, which can be simplified further, as before, by grouping nodes for each day together and calculating contact matrices and distributions.



(a) set1 (b) set2

Figure 4.9: Histogram of the final size of the epidemic for two parameter sets (see Tab. 4.2) and the various static networks. The lines show simulations on the static networks, calculated on the entire data set. The symbols show simulations on the corresponding daily aggregated networks, a sequence of static networks calculated for each of the 7 days of the dataset.

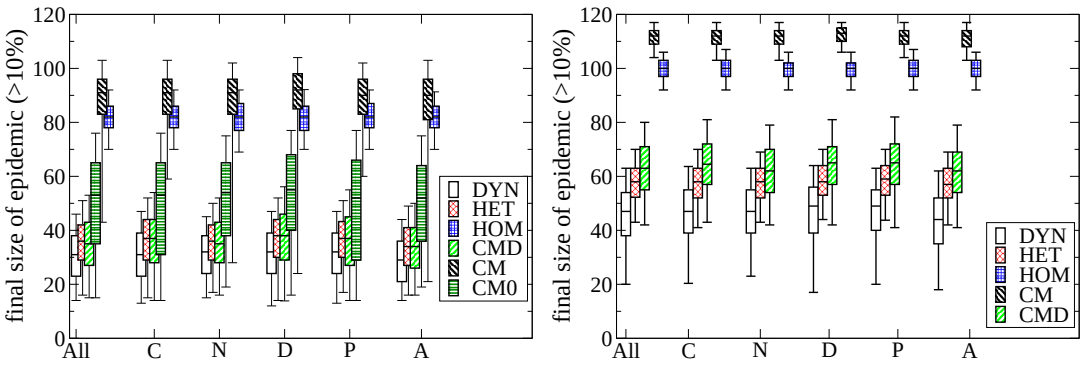
Epidemic spreading was simulated for the networks based on these data representations with different inherent amount of information and the results are shown in Fig. 4.9. Keeping a daily temporal resolution of the data improves the simulation results for all networks. The number of nodes per group varies from day to day. Thus, even though in the contact matrix networks all present nodes are connected with each other for every single day, they are not directly connected to all nodes appearing on other days as well since not all nodes are available for spreading all the time. Only for the sequence of daily CMD networks, the effect of the lower number of daily present nodes on the epidemic spread is counteracted by an increased number of connections on the fully aggregated network. Since links are placed randomly for every daily contact matrix network, mixing among nodes increases slightly compared to the CMD network on the entire data.

4.2.4 Influence of roles

The results of epidemic simulations have been averaged over different starting times and starting seeds. The role of starting times has been assessed in Sec. 3.2. Here we consider the influence of the class of the epidemic seed and how the probability of getting infected depends on a node's class. Due to their inherent properties, different nodes have indeed different risks to get infected during the epidemic. We are in particular interested in the role a node plays for the course of the epidemic depending on the group to which it belongs. As policies for the prevention of epidemics cannot be individual based, the affiliation to a given role class can be used as a proxy in order

to predict the importance of single individuals for epidemic spreading in absence of more precise information.

We find that the seed plays an important role for the extinction probability of the epidemic and also in the first phase of the epidemic. The extinction probabilities and the probabilities for an epidemic to reach less than 10% of the population are given in the appendix in Tab. A.4 and Tab. A.5 for networks based on different data representations and depending on the class to which the seed node belongs. Epidemics starting from an assistant are most likely to spread and reach a larger part of the network, while epidemics starting from a patient, caregiver or doctor are more likely to die out or only reach a small part of the network. The extinction probabilities are qualitatively similar for the HET and CMD network. On the other hand, the probability to die out when spreading starts from a patient or caregiver is underestimated in the CM representation. It is comparable to the case when the seed is a nurse.



(a) set1 (b) set2

Figure 4.10: Boxplot for the final size of the epidemic for epidemics which reach more than 10% of the population, depending on the group to which the starting seed belongs. These groups are: C- Caregivers, N-Nurses, D- Doctors, P-Patients, A-Assistants. The case where starting seeds are chosen randomly from all groups is marked as "All". The boxes show the median, 25% and 75% quantiles, the whiskers show the 5% and 95% quantiles. The parameter sets are given in Tab. 4.2

Once the epidemic has infected many different nodes, the choice of the seed is of no importance anymore for the size of the outbreak of the epidemic. In Fig. 4.10 it can be seen that the outbreak size for outbreaks which reach more than 10% of the population is independent of the group of the starting seed but depends strongly on the data representation used for the simulation.

Nodes which have the highest impact on the epidemic as a starting node are also the most in danger to get infected during an epidemic. In Fig. 4.11 and Fig. 4.12 the fraction of nodes of each group which gets infected during an epidemic is shown. Assistants have a very high chance to get infected during an epidemic, whereas patients and caregivers are fairly safe. The CM representation and the HOM network do not only overestimate the outcome of the epidemic greatly, simulations on these networks also misjudge the respective importance of the different groups. In the CM representation, patients and caregivers appear to be almost as much in danger to become infected by an epidemic as nurses. Due to the high strength and low degree of the patient and caregiver groups, the importance of these groups is overestimated in a data representation which does not retain information on the degree and heterogeneity of the weights.

When conceiving immunization schemes, knowing which groups are most endangered plays a major role. The limitations of the CM model when faced with weight distributions which are

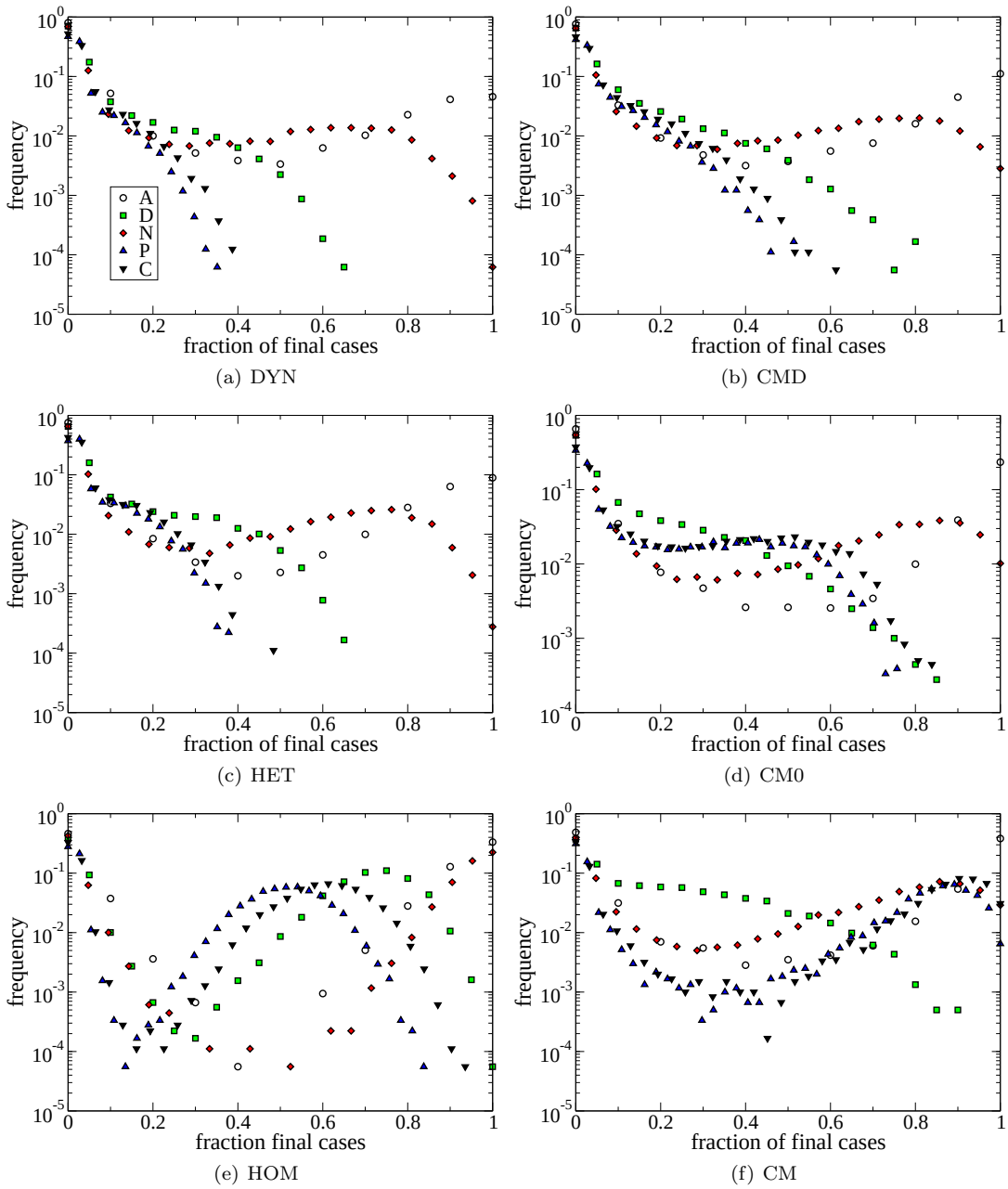
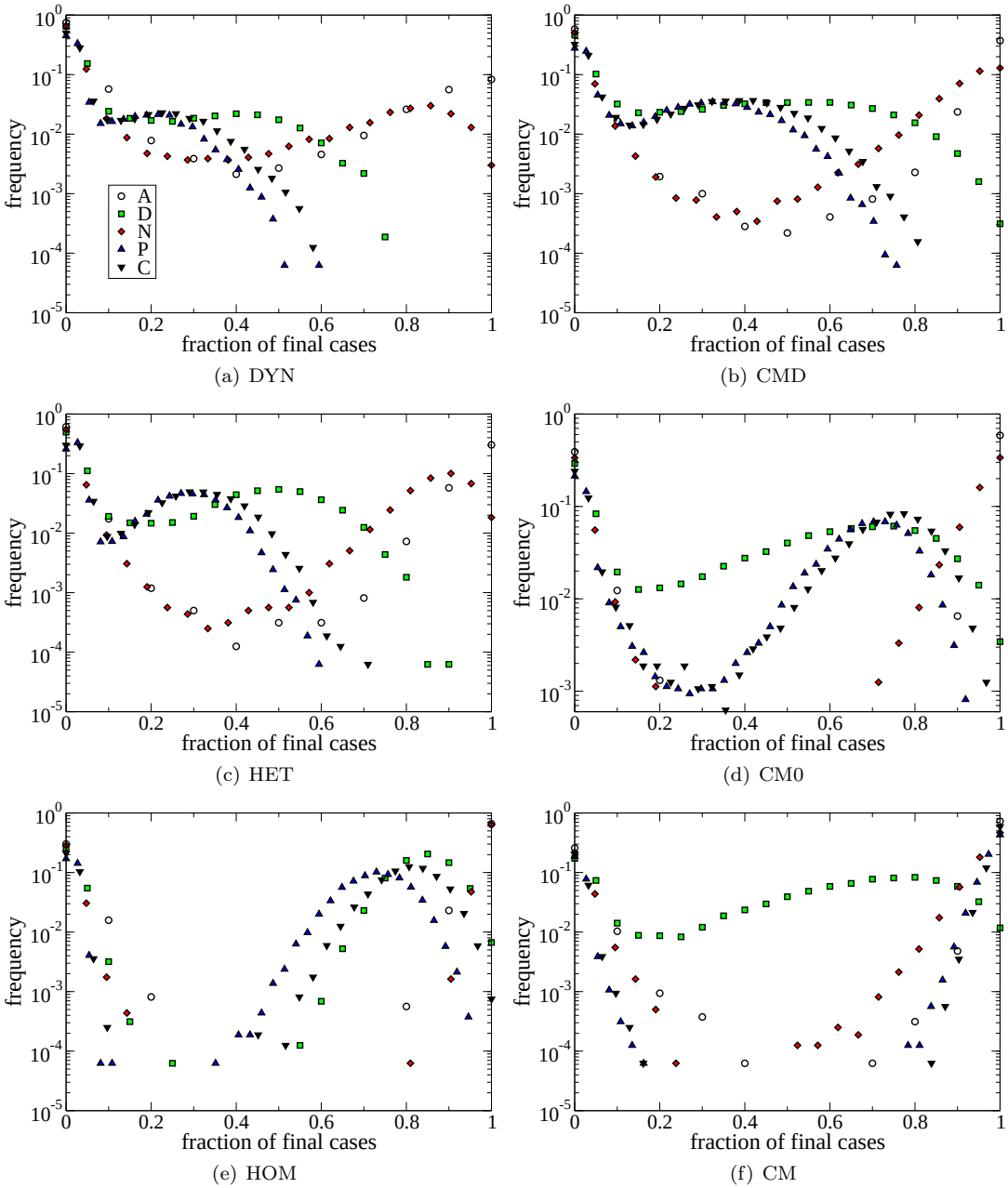


Figure 4.11: Histogram of the fraction of members of each group which are infected in the course of an epidemic outbreak for parameter set 1 (see Tab. 4.2). Simulations are done on the DYN, HET, HOM, CMD, CM0 and CM network.

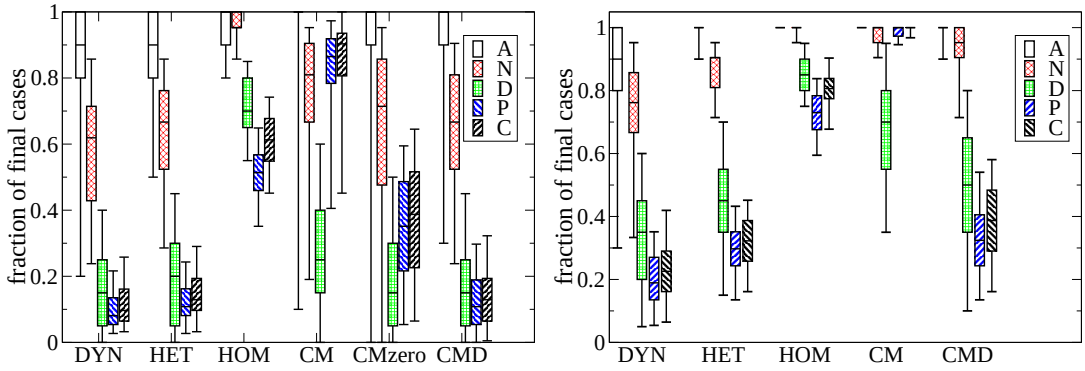


(a) DYN (b) CMD

(c) HET (d) CM0

(e) HOM (f) CM

Figure 4.12: Simulation for parameter set 2, as in Fig. 4.11.

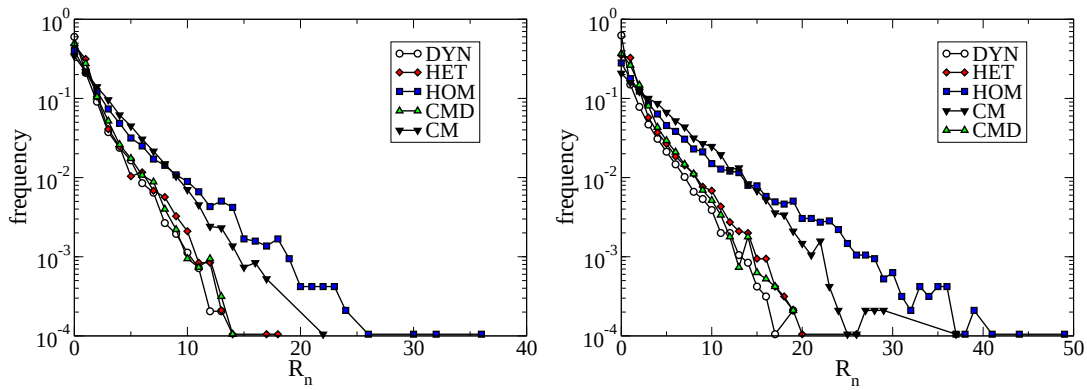


(a) set1 (b) set2

Figure 4.13: Boxplot for the fraction of nodes in each group which were infected during the course of an epidemic outbreak that reached more than 10% of the population. The boxes show the median, 25% and 75% quantiles, the whiskers show the 5% and 95% quantiles. Simulations were done for two parameter sets (see Tab. 4.2).

broad and contain many zeros needs to be taken into account and can be overcome by introducing distributions to the contact matrix.

4.2.5 R_0-correction



(a) set1 (b) set2

Figure 4.14: Distribution of the number of secondary infections R_n per node for the different networks for two parameter sets (see Tab. 4.2). The distribution is based on 80 simulations starting from each of the nodes in the network.

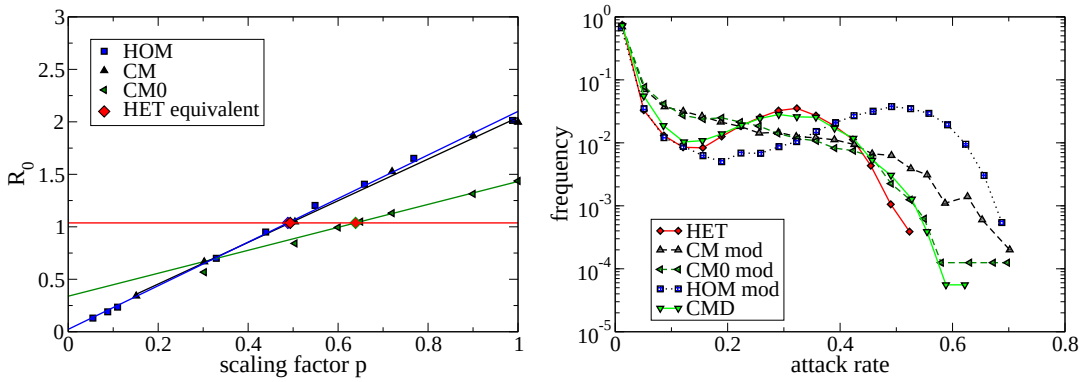
Replacing heterogeneous link weights by their average leads to an overestimation of the epidemic, also because the transmissibility is a concave function. The transmission probability over two links with very different weights is therefore smaller than the transmission probability over two links which have the respective average weight. In order to account for this fact, we will rescale the probability of propagation so that the average number of secondary cases is the same for all networks. In Fig. 4.14 the distribution of the number of secondary cases on the networks is shown for the two parameter sets. The distribution is exponential and it is most likely that the seed does not infect any other node. The distribution for the HOM and CM networks is broader than for the networks with heterogeneous weight distributions. The average number of secondary cases for these networks is higher as well (Tab. 4.4). The average number of secondary cases for the dynamic network DYN is smaller than for the HET network, mainly because the epidemic has a much higher extinction probability on the DYN network. This is due to the burstiness of the network, which leads to comparably longer times on average before a node comes into contact with other nodes. If the epidemic starts with the introduction of the seed into the network, the probability that nodes recover before they have the possibility to infect other nodes disappears, and the average number of secondary cases increases to a value comparable to HET or even higher. In the case of parameter set 2, for example, this more than doubles the average number of secondary cases, from 1.06 to 2.32.

As the average number of secondary cases for the HOM, CM and CM0 networks are much higher than for the HET network, we rescale the spreading probability β for HOM, CM and CM0, so that they have the same average number of secondary cases as HET. In order to find the right scaling parameter, we recalculate the average number of secondary cases for simulations with different scaling parameters in Fig. 4.15(a) and fit a linear function through the data points.

Nevertheless, even after correcting for the average number of secondary cases, HOM still overestimates the outcome of the epidemic (Fig. 4.15(b)). The shape of the distribution of final cases for CM and CM0 remains very different from the one for DYN and HET. Low as well as high outcomes of the epidemic are now overestimated by those data representations. Furthermore, the relative importance of groups is still not assessed correctly (Fig. 4.16).

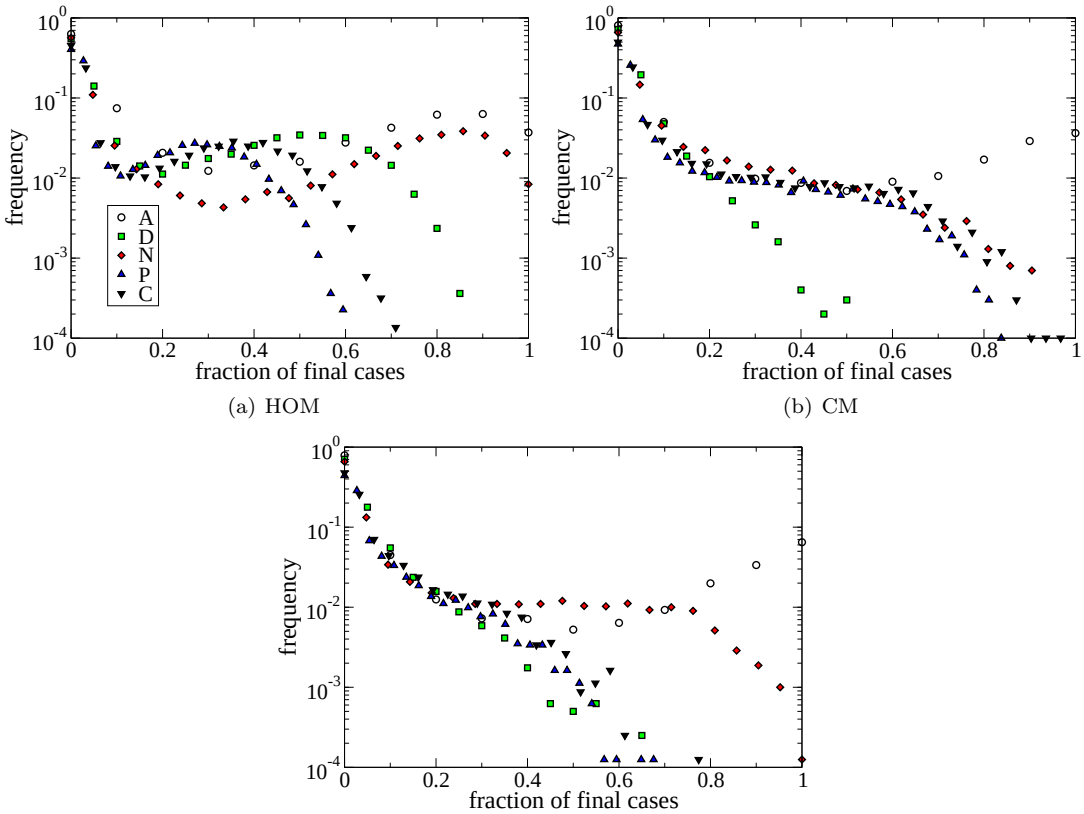
	set1	set2
DYN	0.84	1.06
HET	1.04	1.66
CMD	1.04	1.67
CM0	1.44	2.27
CM	2.00	3.87
HOM	2.06	3.71

Table 4.4: Average number of secondary infections for the different network and two parameter sets (see Tab. 4.2)



(a) (b)

Figure 4.15: (a) Average number of secondary infections for the HOM, CM and CM0 networks as a function of a scaling parameter p, which scales the probability of propagation (b) Distribution of the final size of the epidemics. Simulations for the HOM, CM, and CM0 network are done with modified propagation probabilities so that the average number of secondary infections is identical to the one on the HET network with parameter set 1.



(a) HOM

(b) CM

(c) CM0

Figure 4.16: Fraction of final cases in each group for the networks HOM,CM and CM0 with modified propagation probabilities so that the average number of secondary cases is identical to the one on the HET network with parameter set 1.

4.3 Conclusion

We have contributed to the question, which level of detail of the data is necessary or sufficient for the simulation of epidemic processes on the data. We estimated which resolution in time is necessary for epidemic simulations on dynamic networks depending on the spreading parameters. Furthermore, we have introduced a data representation which bridges the gap between contact matrices and dynamic network data. By including weight distributions to the usual contact matrix representation, we could incorporate the heterogeneity of link weights. This greatly improved the accuracy of simulation results especially in cases where the weight distributions on the original data are very broad. In the next chapter, we will compare the different data representations tested in this chapter with regard to their capability to devise immunization strategies.

Chapter 5

Immunization on dynamic networks and data representations

One of the most important areas of epidemiology is disease prevention. Since the introduction of vaccination, major steps have been done. People who are vaccinated do not only acquire immunity for themselves but also protect society as a whole. When a certain fraction, depending on the reproductive rate of the disease, has been vaccinated, herd immunity can be acquired, effectively stopping the epidemic as a whole. However, as many social networks have a broad degree distribution [52], to acquire herd immunity almost the entire population needs to be vaccinated [77]. On the other hand, scale-free networks are very susceptible to targeted attacks [24, 25, 77], and therefore targeted immunization can effectively reduce the outbreak size by only vaccinating a small portion of the population.

Having data to inform social network models can therefore lead to more efficient immunization policies. An optimal immunization scheme could be reached by general surveillance of the interactions between individuals by means of mobile phone applications, for example through Bluetooth and GPS, as well as self-reported information. However, if this data acquisition stays on a voluntary basis, it will not cover the whole population. Furthermore, it raises severe privacy issues.

So far, on a population level no data for a complete network view exists, and proxies have to be used to guess possible transmission pathways. Therefore, in many epidemiological studies contact matrix models are still widely used [35, 36, 57, 79, 105]. The contact matrix can, for example, be defined by the average contact time or the assumed reproductive rate between age groups or roles. The lack of exact data has a direct influence on the epidemic simulations (see Chap. 4) and also has a direct effect on the realizability of optimal immunization strategies that rely on global network information. The possibility to find the super-spreaders [55] in the population is greatly dependent on accurate data.

Some efforts have been done to find practicable immunization schemes which do not rely on global or complete information. These are for example the method of acquaintance immunization [25]. Here individuals are randomly selected and one of their friends is immunized. This systematically leads to a selection of random individuals with a higher than average degree, as our friends usually have a higher degree [31] than we do. However, due to its random elements, this immunization method remains far from being optimal. It can already be outperformed by

immunization strategies based on only a short amount of collected exact network data [86].

Immunization policies need to make practical immunization suggestions which are general and easy to enforce. Therefore two challenges have to be met to deduce a useful immunization strategy from real data. The strategy needs to be generalizable and transferable to other similar datasets, and it needs to hold also for future events. In the last chapter we have introduced a data representation which is group-based and fairly general without losing too much of its predictive power. Results for epidemic simulations using this data representation were comparable to results using the exact temporal data. Here we will test the capability of this data representation to suggest successful vaccination strategies. In the next chapter (Ch. 6), we will see that there might be limits to predicting the perfect set of nodes to vaccinate even with complete data, as presence and importance of nodes change over time. It might therefore not be the optimal choice to vaccinate a super-spreader due to past evidence as correlation with future performance might not be overly high. On the other hand, common sense tells us that people are different, that their social activity might fluctuate around different averages or following different distributions. Thus, a short dataset might not be sufficient to find an exact ranking between individuals but it might suffice for an approximate one. Also, as belonging to a group might transfer some of the group properties to its members, group-wise vaccination might be a feasible alternative. Finding a proxy, like group-membership, to describe the importance of nodes is a feature that is easily generalized. We are looking for generalizable immunization strategies which can rely on limited data without running the risk of false predictions due to too few data.

Even though complete dynamic data does not exist on a nation wide scale, for smaller settings the differences between common low-information and high-resolution networks and the resulting influence of the information on epidemic prediction and immunization strategies can be tested.

In this section, we will test different immunization strategies on a dataset describing the interactions in a hospital ward. Current immunization strategies, suggested by the Center for Disease Control and Prevention, are to prioritize health-care workers but do not make further distinctions between different health-care workers [79].

Most strategies have been explored on static networks or using contact matrices, where no change of individual importance for the disease arises. We test the effect of the data representation on the prediction of efficiency of different simple group-based immunization strategies in Sec. 5.1. In Sec. 5.2 we compare the efficiency of immunization strategies based on different data representations. For more general data representations, individual features of nodes disappear and only group information is retained. Due to higher generality, better enforceability and less privacy issues, group based immunization strategies appear to be a very practical alternative to randomized strategies or to strategies that rely on exact network data. In Sec.5.3 we focus our comparison between group based immunization strategies and individual based immunization strategies on the stability of these strategies over time and also on their dependence on the length of the dataset. In Sec. 5.4 we control for the starting-time dependence of immunization strategies based on daily networks in order to test the reliability of these strategies for future events. In Sec 5.5 we will look into possible methods to find an optimized vaccination strategy that also makes use of the time-resolved information of dynamic networks and discuss their limits.

5.1 Influence of the data representations

To evaluate the efficiency of an immunization strategy via simulations, one has to choose a model for the epidemic, a data representation on which to run the model and a measure for the outcome of the simulation in order to evaluate and compare the different strategies. We

use the SEIR model for simulations as before. Our focus here is on the influence of different data representations on the evaluation of the immunization strategy. In Chap. 4 we have seen that the data representation does have an influence on the outcome of epidemic simulations. Here we will test whether there is a difference in predicting the efficiency of vaccination for different data representations and if it is crucial. We use a simple immunization scheme, in which ten random individuals from one of the five groups (assistants, doctors, nurses, patients and caregivers) are vaccinated. The same five data representations as before (DYN, HET, HOM, CMD, CM) are used for simulations. The distributions of the final size of the epidemic simulated on the different data representations can then be compared (see Fig. 5.2 and 5.1). The case without immunization is added as a reference for all five data representations. For all five data representations, a correspondence can be seen between the groups which are the most likely to be attained by the epidemic and the groups which, when vaccinated, lead to the highest reduction on the outcome of the epidemic. For the DYN, HET and CMD data representation, the best groups to vaccinate are assistants and nurses. Vaccinating patients and caregivers is the least efficient. Even though the overall case counts for epidemics on the HOM data representation are greatly overestimated, and the effect of vaccination are quantitatively different, the order of importance between the different groups is the same as for the DYN network. Only the CM representation grossly overestimates the efficiency of vaccinating the patient and caregiver groups, just as it overestimated before their probability to catch the disease. Simulations on this data representation can therefore lead to misleading conclusions about the right group to vaccinate, with possibly severe effects. It is also noticeable that vaccinating patients and caregivers does not reduce the attack rate. Vaccinating these groups does not have a strong effect as group members have a low probability of getting infected in the course of an epidemic (see Sec. 4.2.4).

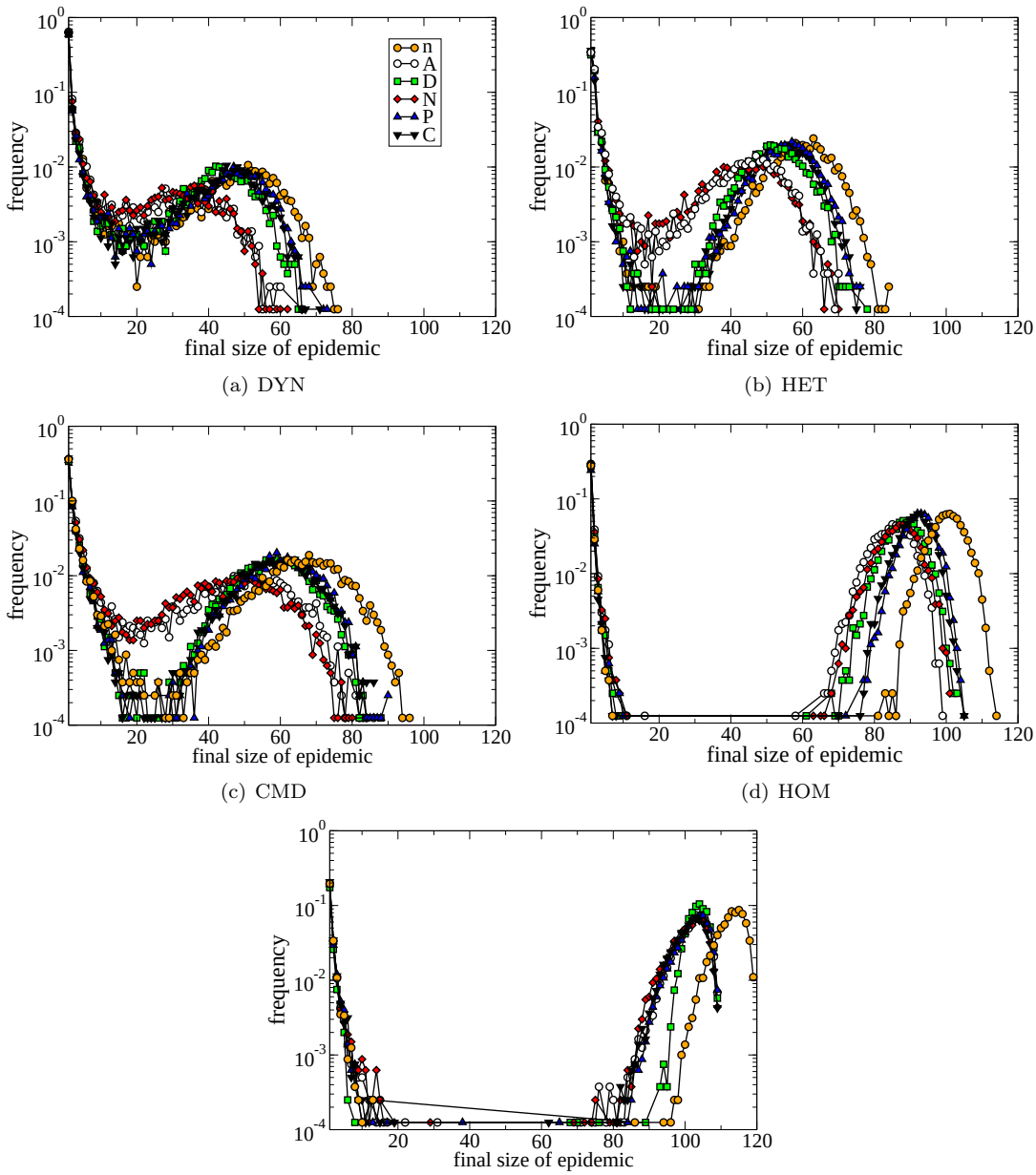
5.2 Immunization strategies on static data representations

Here we evaluate the efficiency of the optimal immunization strategies that can be derived from each data representation. As strategies are chosen to be optimal on each data representation, it makes little sense to run the simulations on networks created from the corresponding data representation. In order to compare the different immunization strategies, simulations will have to be run on the same underlying data representation. As the DYN representation of the data has the highest amount of information and is the most similar to reality, we choose this representation as our gold standard. Simulations are run using the same SEIR model for simulations as before. Comparison between the different immunization strategies will be done by considering the percentage of simulations with an attack rate below 10% and the median final number of cases for simulations which result in an attack rate above 10%. The immunization strategy is more efficient when more runs have an attack rate below 10% and the median number of cases is low.

As the different data representations have different inherent amounts of information, vaccination by individual degree of nodes is not available for all data representations. Of the data representations used here, only DYN, HET and HOM have individual information for all nodes, and therefore only those three data representations can provide enough information for degree based immunization schemes.

The CM and CMD representation do not contain other information on individual nodes than their belonging to different classes with different class properties. Only class based immunization schemes can be formed based on these data representations.

In the CM representation, classes can be ordered according to the total time individuals of each class spent in contact. Inside classes the immunization ranking has to remain random, as



(a) DYN

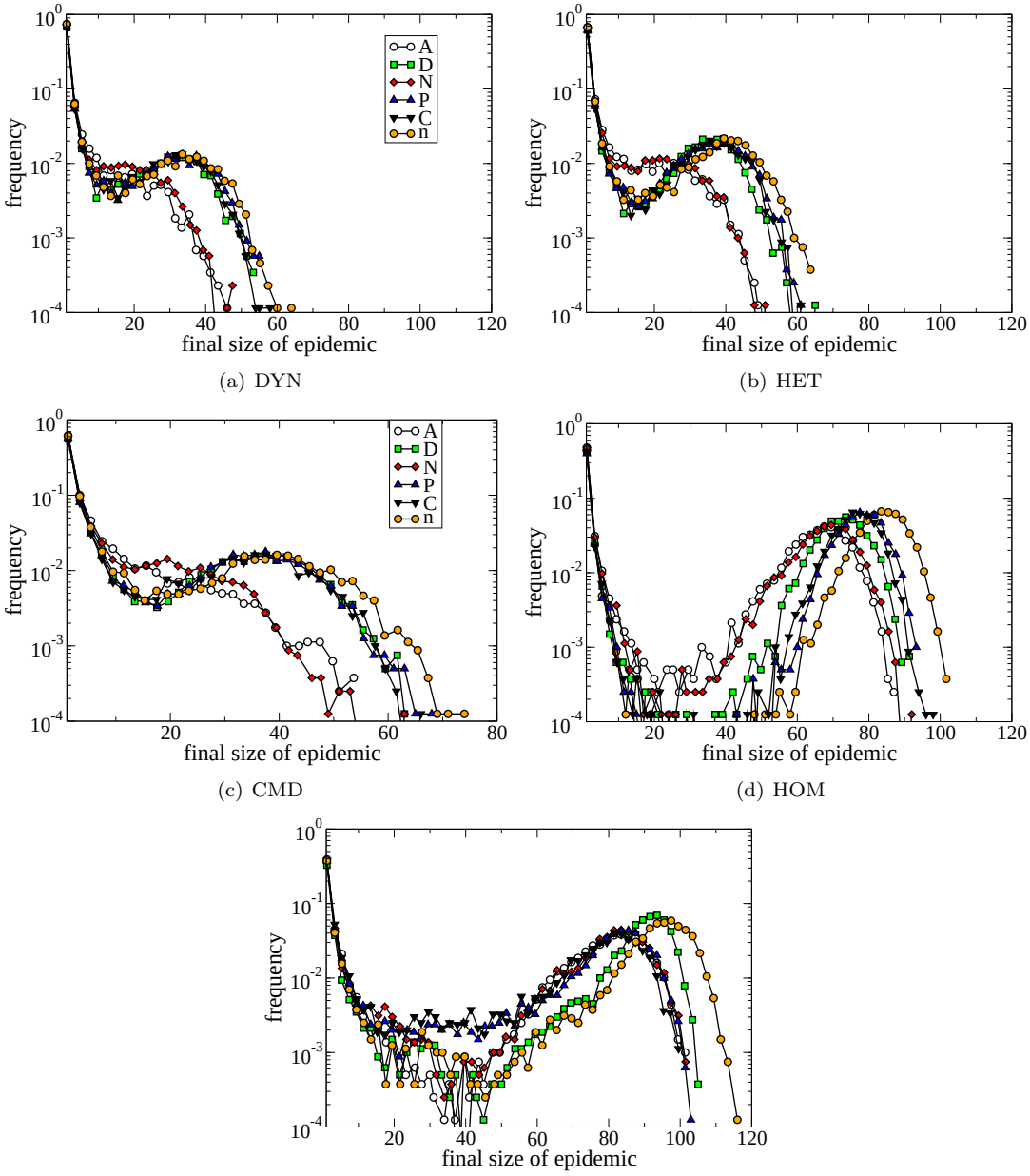
(b) HET

(c) CMD

(d) HOM

(e) CM

Figure 5.1: Vaccination of 10 subjects out of one of the groups (A,N,D,P,C) or no vaccination at all (n). Simulations were done with parameter set 2 (see Tab. 4.2) on the various networks.



(a) DYN

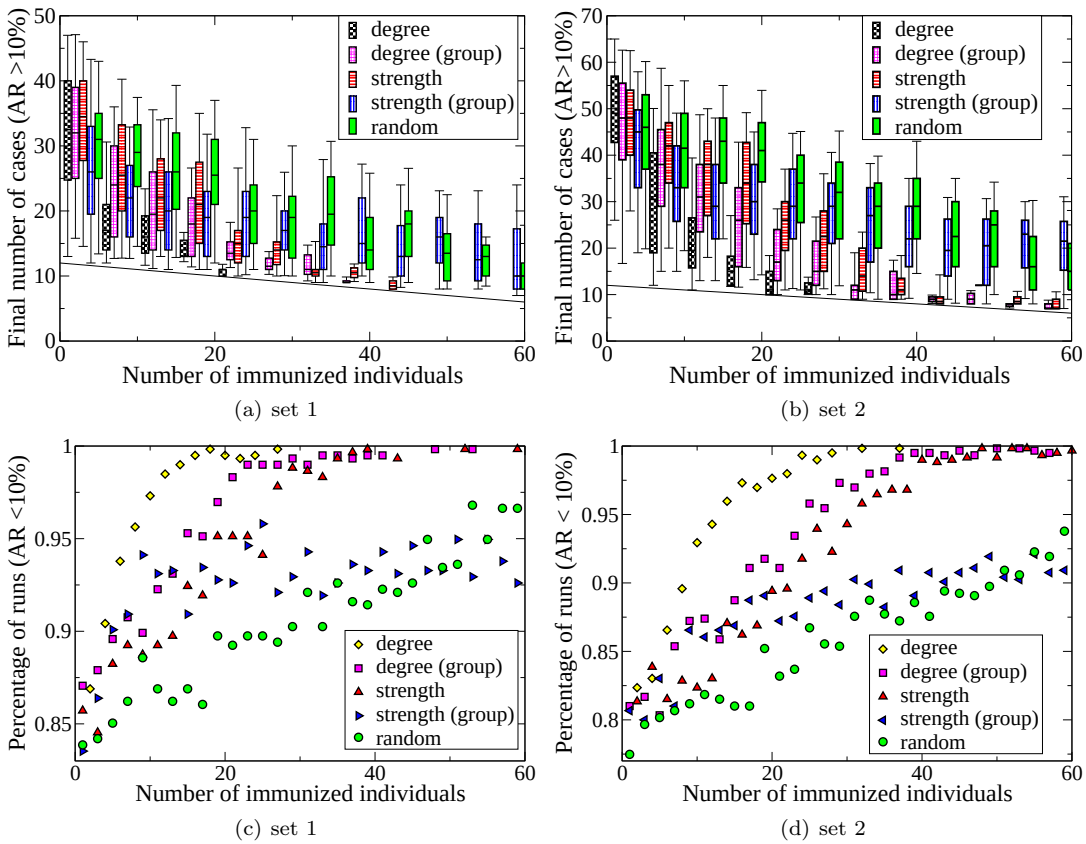
(b) HET

(c) CMD

(d) HOM

(e) CM

Figure 5.2: Vaccination of 10 subjects out of one of the groups (A,N,D,P,C) or no vaccination at all (n). Simulations were done with parameter set 1 (see Tab. 4.2) on the various networks.



(a) set 1 (b) set 2

(c) set 1 (d) set 2

Figure 5.3: Effect of the immunization of an increasing number of participants, following different strategies. The left figures are simulated with parameter set 1, the right figures with parameter set 2 (see Tab. 4.2). The top figures show boxplots of the final size of epidemics, restricted to the runs in which more than 10% of the non-immunized population was affected. The gray line marks the 10%. The box represents the median and the 25% and 75% quantiles, the whiskers the 5% and 95% quantiles. The bottom figures show the fraction of simulations that yield an attack rate smaller than 10% of the non-immunized population. The immunization strategies are: "degree", which immunizes individuals according to their degree on the aggregated static network, immunizing highest-degree nodes first. "degree (group)" immunizes individuals from groups with the highest average degree first, the ranking within the group is random. "strength" immunizes individuals according to their strength on the static network, "strength(groups)" ranks groups according to their average strength, individuals within each group are ranked at random. "random" is a complete random choice of individuals for immunization.

no information to discriminate between nodes is available.

In the CMD representation, in addition to the information of total contact time, we also have information of the link density within and among classes. This allows a ranking of classes according to their average degree.

As a reference we also test the random strategy, which is the only possible strategy if no information is available to distinguish individuals. This is for example the case for the fully connected networks with identical link weights, describing the homogeneous mixing case.

In Fig. 5.2, plots are shown for both parameter sets. Vaccination by degree performs best, followed by the strategy based on the CMD data representation. Vaccination by average strength of groups partly performs worse than the random strategy. This is due to the fact, that the patient and caregiver groups include individuals with high contact rates but very low degree. This leads to a high average strength for these groups but low average degree. Often, strength and degree are correlated so that groups with high average strength are also groups with high average degree. This is not the case here (see Fig. 2.2). Therefore the information on strength of groups cannot serve as a proxy for the degree of groups. As the CM representation does not retain information on the degree of groups, it can only inform on the vaccination by strength of groups. Vaccination by strength for these groups proves to be inefficient.

5.3 Effect of a limited time window

The data representations also have a limit in the available information that is imposed by the length of the dataset. In this case, the dataset extends over one week.

The ranking could differ for different periods of data acquisition. Here we can only subsample within the period given by the dataset. We will therefore reduce the data to one-day samples and to half-day samples in order to investigate the effect of the time of data acquisition and the limit of information imposed by the dataset length on data-based immunization strategies. In Fig. 5.4 and 5.5 we compare individual vaccination by degree, based on the data aggregated on one day only, for different days, with class vaccination by degree where the information is also limited to one day. For four out of seven days, the group-based immunization scheme stays the same as the one calculated on the complete dataset. For the other 3 days, the ranking still considers assistants and nurses as the most important classes to vaccinate. The respective vaccination orders are NADCP and NACDP. The efficiency of these three group-based immunization schemes is very similar. The efficiency of the individual-based immunization schemes decreases slightly when information is reduced to only one day of contacts.

Ranking on half days leads again to similar results, classing A and N first in 13 out of 15 cases, C and P last, when groups were ranked according to their degree. If only the information of the CM representation is available, and groups cannot be ranked according to their average degree, then the Patient and Caregiver groups are very often classed as the most important groups as they have a high strength on the daily contact matrices.

Nodes which have a high degree on one day do not necessarily also have a high degree on other days. As we will see in Sec. 6.1, the ranking by degree fluctuates in time. In Fig. 6.7 we see the development of the degree ranking over time for the diverse roles. The ranking of different roles, especially nurses in comparison to patients, shows a rather stable and distinct behavior, while the individual ranking fluctuates. The precise ranking according to individual degree is therefore less stable than the ranking according to group degree. The individual degree ranking loses some of its advantage when it is applied to parts of the dataset on which it has not been constructed. Thus the difference between the efficiency of the immunization scheme based on the DYN network and the efficiency of the immunization scheme based on the CMD network

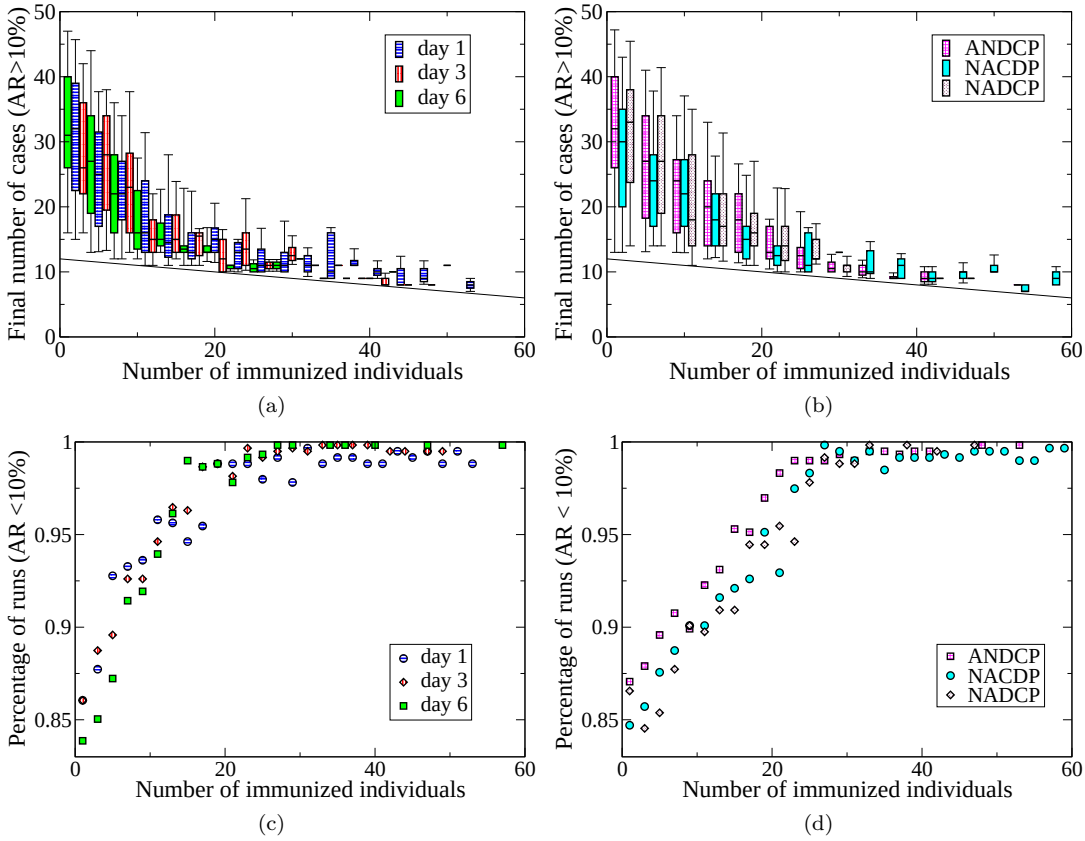


Figure 5.4: Effect of the immunization of an increasing number of participants. In the left figures (a+c), the immunization strategy is based on the degree of the individuals calculated on the aggregated network of limited length of one day. In the right figures (b+d), the immunization strategy is "degree (group)", where the average degree of the classes is calculated and ranked on daily aggregated networks. On the 7 days, 3 different rankings of groups are found and shown here. Individuals within each group are ranked randomly. The top figures (a+b) show boxplots of the final size of epidemics, restricted to the runs in which more than 10% of the non-immunized population was affected. The box marks the median and the 25% and 75% quantiles, the whiskers the 5% and 95% quantiles. The bottom figures (c+d) show the fraction of simulations that yield an attack rate smaller than 10% of the non-immunized population. Simulations were done on the dynamic network with parameter set 1.

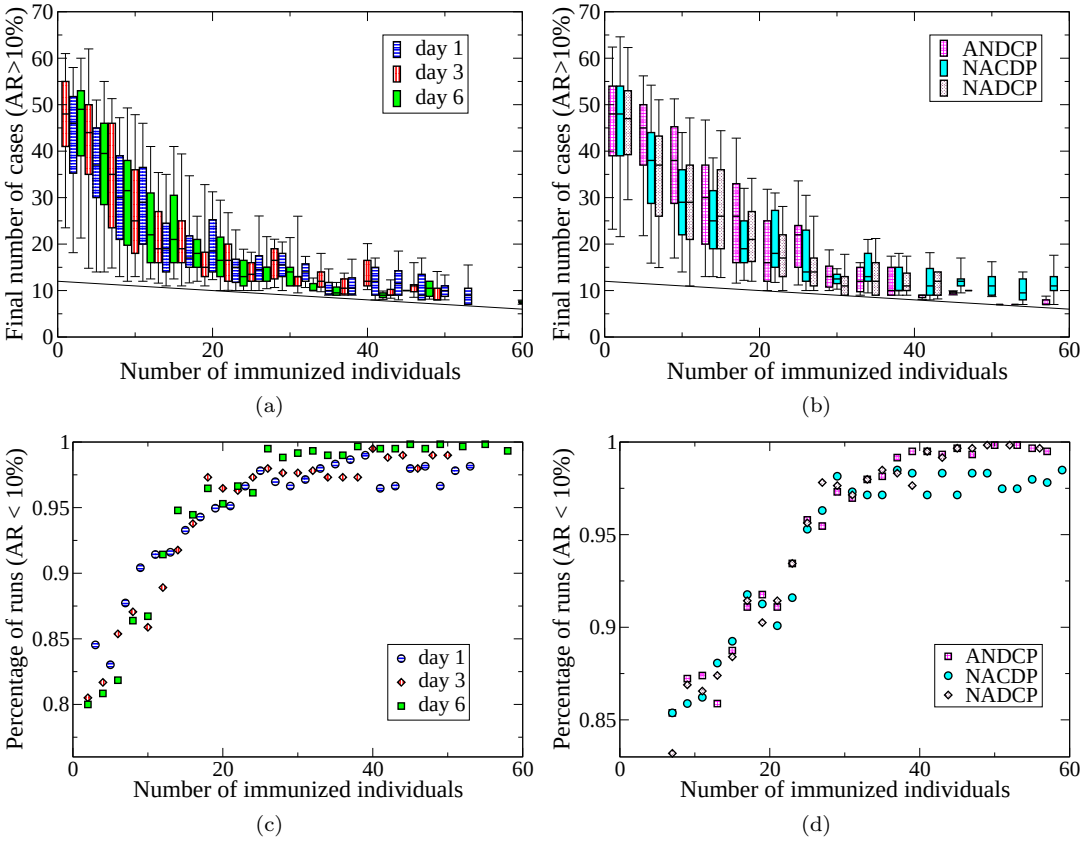


Figure 5.5: Same as Fig. 5.4. Simulations were done on the dynamic network with parameter set 2.

decreases slightly for shorter collection times of information.

5.4 Time dependence of ranking efficiency

As the ranking of nodes changes over time, the question arises how valid a ranking chosen at a specific point in time remains in the future. In our simulations, the choice of nodes to vaccinate was taken on the same dynamic network on which the efficiency of the vaccination was subsequently tested. This introduces a bias which increases the immunization efficiency. To explore the size of this bias, we check the efficiency of an immunization scheme depending on the starting time relative to the time at which the scheme was devised.

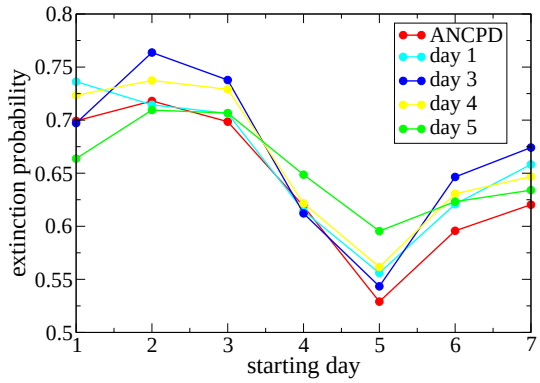


Figure 5.6: The extinction probability is plotted for various immunization schemes as a function of the starting time of the epidemic. The immunization schemes are based on individual degree ranking on data aggregated over one day. Plots for the different days are marked with different colors. In each run, 20 individuals were vaccinated. Lines are guides to the eye.

For daily degree-based rankings on different days, we simulate epidemic processes with different starting times and plot the extinction probability as a function of the day on which the epidemic started in Fig. 5.6. The simulation was done with parameter set 2 (see Tab. 4.2). Epidemics for which the seed node does not infect any of its neighbors have an average duration of one day for this parameter set. Thus, in most cases the seed does not stay infectious over the whole week and the starting time is characteristic for the effect the vaccination has on epidemics which occur at and shortly after the starting time. A slight effect on the dependence of the starting time can be seen. For example, the extinction probability of epidemics starting at the first day is highest for the immunization scheme which was designed on a degree ranking based on the first day, whereas the extinction probability of epidemics starting at the fifth day is highest, if the immunization scheme is based on the degree ranking of nodes on day 5. Nevertheless, the overall dependence of the extinction probability on the starting time due to weekly fluctuations of the data is much higher.

In order to focus on the effect of the dependence of the outcome on the relation between the starting time and the day on which the vaccination scheme was based, and in order to factor out the effect of the starting time due to variations of the data, in Fig. 5.7 we have plotted the ratio of the extinction probability with and without immunization for two vaccination schemes based on the degree ranking at different days. This ratio is lowest when the immunization is not very efficient and highest when the immunization strategy is very efficient. We can see that it is most efficient for epidemics which start at or just before the day on which the ranking was chosen and

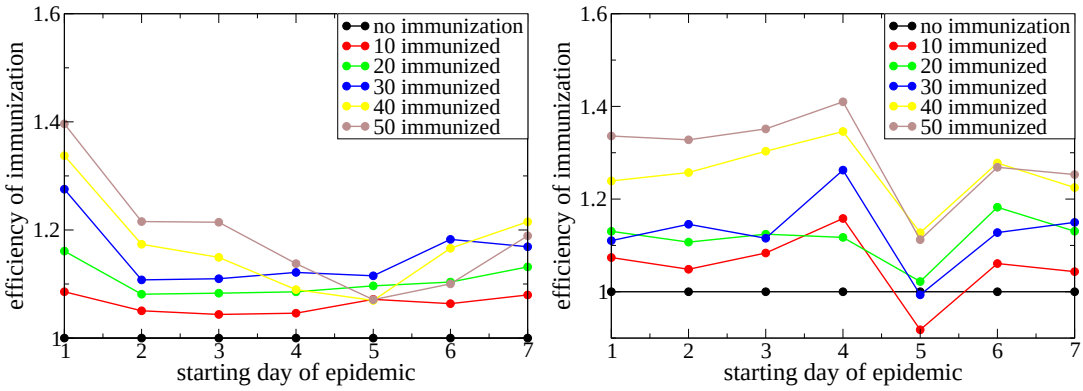


Figure 5.7: The ratio of the extinction probability for epidemics with immunization to the extinction probability for epidemics without immunization is plotted against the starting time for different numbers of vaccinated individuals. (a) The immunization ranking was based on the aggregated network of day 1. (b) The immunization ranking was based on the aggregated network of day 4.

worst for epidemics which start just afterwards, as due to the cyclic nature of the network, the epidemic is then least likely to reach the day which the immunization strategy is based on.

However, these effects are very weak and more simulations have to be done in order to see if these results are significant or not. Especially the exact choice of the time window over which the immunization scheme is calculated can have an effect. Here we have chosen 24 hour time windows. Due to daily repeated fluctuation patterns, this is a reasonable choice. However, time windows do not start at midnight but at the beginning of the dataset in order to maximize their number. This could possibly have a small influence on the outcome. It remains to be seen how much the daily-based immunization schemes need to differ when based on successive days so that their efficiency changes strongly with the starting time of the epidemic.

5.5 Immunization strategies on dynamic networks: significance

Dynamic networks differ from static networks through the presence of dynamic motifs [4, 61]. The spreading between two nodes which are neighbors in the static network can be impossible in the dynamic network if their contact takes place before either of them gets infected.

It could be important to take dynamic aspects into account when devising immunization methods. So far, methods have been put forward which easily take account of the different occurrence probabilities of nodes by vaccinating the most recent or the most frequent contact of a random node [51]. However, they do not take account of network motifs or of the importance of single nodes for the efficiency of spreading paths. The advantage of these strategies is that they are easy to apply and better than complete random strategies. However, they still include random elements and are not optimal strategies. Understanding better which characteristics lead to an optimal immunization strategy could also improve these more practicable methods. As many people know their friends quite well, being able to point out characteristics of efficient spreaders could then improve simple strategies in which randomly chosen subjects point out the

one friend which fits these characteristics best. When the optimal immunization strategy is found, it remains to be seen how much of the efficiency of a spreader is due to its own characteristics and how much to the interactions on the network, the global network structure or just plain chance.

Takaguchi has shown that single events can play an important role in the dynamic network [95]. If these events are removed the dynamic network will fall apart. We explore a new method to evaluate the importance of single nodes in the spreading process.

To this end, we measure the effect the removal of a single node has on temporal paths starting at a specific time. A temporal path between nodes i and j is the fastest path that can be traced between the two nodes on the dynamic network [73]. There can be more than one temporal path, and in that case, we choose the one which includes fewer nodes. There may be no temporal path between two nodes even if they are connected on the static network [73]. For every node in the dynamic network we look at its temporal out-component [96].

The temporal path can be characterized by two properties: the time it took for information from node i to reach node j and the number of nodes that had to transmit this information. We look at the change of both properties when a specific node n in the network is removed. The more significant the position of this node was, the more the temporal paths in the network should change. We therefore quantify the effect the removal of a node has on temporal paths on the dynamic network and call it the node's significance. We hereby distinguish between a change in path length and a change in the duration of the temporal path. This duration of the temporal path from its start at node i up to its arrival at node j can be taken entirely or only measured when node n is in contact. Using this internal clock time of a node has proven to lead to more robust results when considering the time a temporal path takes to reach different nodes in the network [74].

Algorithm to calculate the significance:

- Divide network into different time slices ΔT.
- Choose the size of each slice in such a way that all slices contain the same total contact time between all active nodes.
- For all nodes in one slice, start temporal paths at the first occurrence of each node s.
- Register the time t_{sl} and the number of steps d_{sl} between the seed node s and any node l that it can reach on the temporal network.
- We considered three different ways to measure the time t between the seed and the leaves of the temporal out-component.
  - The total time that has passed, t_T.
  - The total time that has passed and in which the leave node was in contact, its internal clock time or the time it was present, t_p.
  - The total time in which other nodes were in contact with the leave node between the start of the information flow and the time in which the leave node is reached, t_c.
- For each node n which lies on one of these temporal paths, neither as seed nor as leave,
  - remove the node

- register the time t_{sl}^{-n} and number of steps d_{sl}^{-n} for each temporal path from s to l as before.

- The step significance $S_d^{n,\Delta T}$ for node n is the sum over the inverse step difference for all paths.

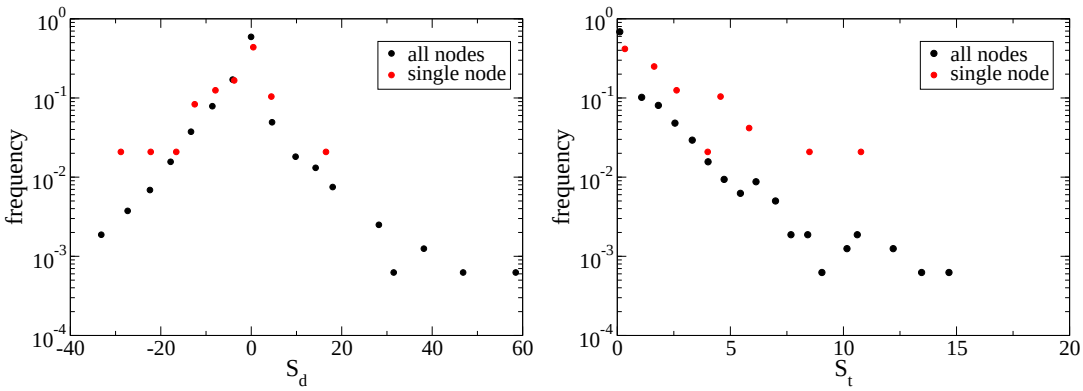
$$S_d^{n,\Delta T} = \sum_{s,l} \frac{1}{d_{sl}} - \frac{1}{d_{sl}^{-n}}$$

- The time significance $S_t^{n,\Delta T}$ for the removed node n is the sum over the difference of the inverse times for all paths it lies on, where t can be t_T,t_p or t_c.

$$S_t^{n,\Delta T} = \sum_{s,l} \frac{1}{t_{sl}} - \frac{1}{t_{sl}^{-n}}$$

Whenever a node does not appear in any of the temporal out-components starting at a specific time window, its significance for this time window is zero. We take the difference of inverse time and the difference of inverse distance in order to avoid adding infinite distances or times if nodes are not reached anymore by the temporal path.

The significance S_{t_c} is highly correlated with S_{t_p}, mostly because nodes in our networks are rarely in contact with more than one node at the same time, so that t_c and t_p do not differ significantly. We therefore drop time measure t_c.



(a) step significance (b) time significance

Figure 5.8: (a) distribution of the significance $S_d^{n,\Delta T}$, (b) distribution of the significance $S_t^{n,\Delta T}$ calculated for the internal time t_p over the complete 4 days of the "lyon2011" data for all nodes and for the most frequently present node.

In Fig. 5.8(b) the distribution of the significances in time, $S_t^{n,\Delta T}$, is shown, and in Fig. 5.8(a) we show the distribution of the step significances, $S_d^{n,\Delta T}$. Simulations were done on the "lyon2011" dataset. Since the temporal path is the fastest path between two nodes, the time of the temporal path between these two nodes can only increase or stay the same when one of the nodes that is lying on this temporal path is removed. Therefore $S_d^{t,\Delta T}$ is always positive. The number of steps between the two nodes, on the other hand, can increase or decrease when

one node on the path is removed. The step significance $S_d^{n,\Delta T}$ can be negative or positive. The significance in time is distributed exponentially. For most temporal paths between two nodes, removing a particular node does not have any significant influence on the temporal path. If this is the case for all temporal paths starting in a specific time window ΔT, then the significance $S^{n,\Delta T}$ of this node is zero, or close to zero. While most of the time the removal of a particular node does not play an important role, sometimes it does.

The importance between nodes and over times varies very much. Looking at the distribution of the significance in time for just one single node, the same variation of the significance can be seen. The shape of the distribution of the significance at different times for all nodes is also due to the distribution of the significance for each single node at different times. For the step significance as well, the distribution for a single node shows the same shape as the overall distribution.

To optimally contain an epidemic it would be sufficient to remove nodes only at those times at which they are most significant. This could, for example, correspond to wearing face masks or other protections in order to not spread diseases at situations when one is in contact with many different people. However, vaccination is not limited in time at the timescales considered here. If we are looking for an immunization strategy based on the temporal information of the network, we need to globally classify and rank nodes. We do this by averaging over the different values at different starting-time windows ΔT_i and then ranking nodes according to their significance.

$$S^n = \frac{1}{N} \sum_{i=0}^{N} S^{n,\Delta T_i} \tag{5.1}$$

Taking the average over the different significance values of the same node at different times reduces much of the information given by this measure. However, if we want to devise an immunization strategy, it does not matter at what time a specific node was important, it only matters if it is likely to be important again. If in the dynamic network some motifs are repeated in contact patterns, they could be detected by our method but not by devising an ordering scheme on a static network.

To distinguish the effect of the significance from the mere presence and absence of nodes, we also include a vaccination ordering in which nodes are ranked according to the number of time windows in which they appear (presence). Similar to the temporal betweenness centrality, we furthermore calculate for each node the percentage of temporal paths (starting at different time windows) in which it is present, independent of the effect its removal would have. This measure should be highly correlated with the temporal betweenness centrality if the number of temporal paths between two nodes i and j, starting at a given time t, is low. Since we only consider one temporal path, the temporal path with the lowest number of intermediate nodes, between any two nodes for the significance, we do the same for this measure, which we call simple temporal betweenness.

We then rank nodes according to the above mentioned classifications of the importance of nodes, their significance S^n, their simple temporal betweenness and their presence and vaccinate them following the respective rankings. This way, we can compare the usefulness of these dynamic classifications with rankings based on information taken on static networks.

In Fig. 5.9(a) we look at the correlation between the significance in time and the degree of nodes. As the degree ranking is quite good as immunization scheme, it serves as a first test for the new significance measure. In Fig. 5.9(b) we compare the significance in time with a more simple but also dynamic measure, the simple temporal betweenness. Both measures are correlated with the significance S_t but the correlation is not so strong as to consider the new measure as redundant.

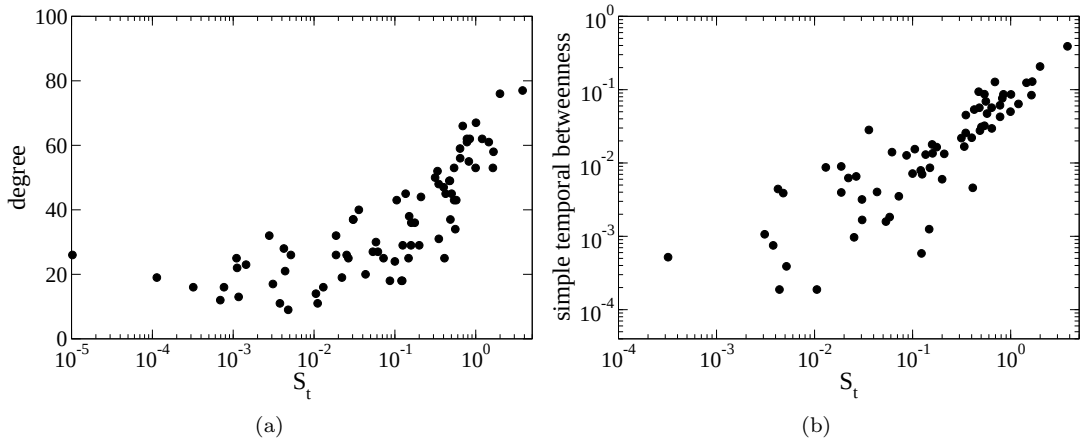


Figure 5.9: (a) For all nodes the average significance S_t is plotted against their degree on the total aggregated network. (b) For all nodes the average significance S_t is plotted against their average simple temporal betweenness.

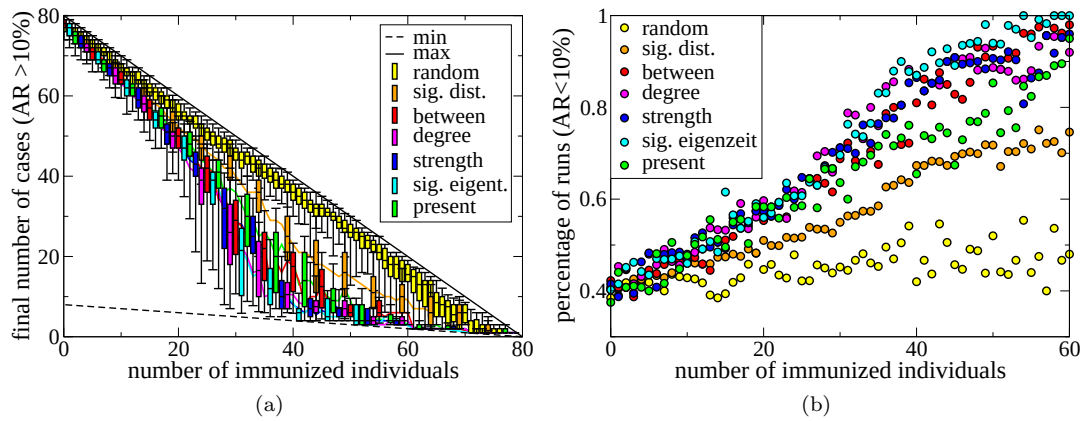


Figure 5.10: Effect of the immunization of an increasing number of participants, following the different strategies (degree, strength, significance S_t, presence). The left plot gives the boxplot of the final size of epidemics, restricted to the runs in which more than 10% of the non-immunized population was affected. The right plot gives the fraction of simulations that yield an attack rate smaller than 10% of the non-immunized population. The parameter set used was: $\beta = 1000 * \mu$,$\mu = 1\text{day}^{-1}$, $\lambda = 1/2\text{day}^{-1}$.

We will subsequently test in Fig. 5.10 if our ranking works better for some networks or some spreading parameter sets than the simpler static and dynamic measures. As the dynamic network features become more important for faster spreading (see Ch. 3), we test the efficiency of the vaccination rankings using epidemic spreading in this parameter region first. The used parameter set is: $\beta = 1000\text{day}^{-1}$, $\mu = 1\text{day}^{-1}$, $\lambda = 0.5\text{day}^{-1}$.

Vaccinating nodes according to the degree ranking has the same effect as vaccination according to the significance ranking. Also, in contrast to the dataset of the children's hospital, vaccination by strength still works quite well. This might be due to the fact that strength and degree are more strongly correlated here. The simple temporal betweenness ranking, which is not based on the effect the removal of a node has and only considers its presence in the temporal paths between nodes, works less well. Finally, the presence ranking, which uses only the information whether or not nodes are present in time windows, is even less efficient. The ranking according to the step significance does poorly, but is still more efficient than random vaccination.

As the significance values are quite widely spread for each node, averaging over those values seems to take out most of the extra temporal information. The remaining information which is inherent in the ranking does not result in better containment of the epidemic than the information already included in the degree ranking, the best among the rankings on static network we consider. Furthermore, the significance only considers the effect of node removal on temporal paths. For epidemics which propagate slowly, the temporal path between two nodes is most likely not the path over which the epidemic spreads. The number of longer paths can play a much more important role, similar to spreading on static networks, where epidemics do not always spread over the shortest path [71]. In Ch. 7 we look into temporal paths and infection paths.

5.6 Conclusion

We have seen that the evaluation of the same immunization strategies on different data representations can come to different conclusions. Using the right data representation to simulate epidemics and predict the efficiency of immunization strategies is therefore very important. In order to devise good immunization strategies, a certain level of detail is necessary. By comparing the efficiency of immunization strategies which could be derived from the information in the different data representations we found that even though individual information on the degree of nodes lead to the best containment of epidemics, information on the average degree as can be extracted from the CMD data representation leads to similar results. Furthermore, when the time over which the information is collected is reduced, the individual degree ranking of nodes became less stable, while group based rankings remained fairly robust. Thus individual ranking schemes need to be considered with a grain of salt. The role single nodes play can change. This could also be seen for the significance of nodes, which varies quite strongly over time. In order to devise immunization strategies, it is important to predict if single nodes or groups of nodes will be important for the epidemic spread in the future. However, any strategy will be based on data from the past. Knowing how much node rankings for immunization strategies change, how much the importance of nodes varies and also, how much datasets vary over time can be an important information in order to evaluate the validity of immunization schemes. In the next chapter we will look at the development of the degree ranking over time, how it changes with longer aggregation time and how it varies over time. We will also compare two datasets which were registered at different times at the same place in order to discern to what extent the datasets differ and to assess the validity of epidemic predictions.

Chapter 6

Predictability

One of the key reasons to make models is not only to understand the main influences of different aspects of reality better but also to make predictions about the future and to learn how to manipulate reality in order to achieve a desired outcome. In order to apply models to the real world, to be able to make predictions, they need to be informed by data. The more specific the data, the more exact are the predictions the model can make. However, if the data is unknown, it needs to be approximated by existing and similar data. The less is known about the underlying situation, the more general the data which is used needs to be.

In this chapter, we will look at the predictability of outcomes concerning the epidemic spreading on networks when generalized data representations are used and also at the stability of rankings of nodes. The data-based rankings can be used as prediction of which nodes going to be the most important in the network.

6.1 Degree ranking

In order to manipulate the course of an epidemic through vaccination, the most influential spreaders need to be known. Similarly to the prediction of the outcome of an epidemic, this task is twofold: knowing the most influential spreaders of a given network and predicting how similar the network will be in the future.

High-degree nodes have been shown to play a crucial role in spreading processes on static networks. In static networks, the ranking of nodes according to their degree centrality is a well-established and straightforward procedure. In dynamic networks, the degree of a node is less well defined. The instantaneous degree of a node, the number of people a person is in contact with at one specific moment, is usually very low, rarely higher than four or five and thus does not differ enough between individual nodes in order to be a good basis for a ranking. The number of distinct nodes a node was in contact with over a longer time period is more meaningful. However, the optimal aggregation time in order to obtain a degree ranking remains unknown. While too short an aggregation time will miss sensible information, aggregating the network over a very long period is costly and the additional data acquisition effort might not bring useful extra information. On the contrary, a prolonged aggregation time might lead to the inclusion of nodes which are not relevant anymore in the ranking list. Furthermore no information about the frequency of occurrence of the links is retained. With longer aggregation more rare links are included in the aggregated network. These links have the same influence on the degree of a node as links with high activity. Nonetheless, the degree centrality is still one of the best and simplest measures to designate influential spreaders, even for dynamic networks [86]. In the last

chapter we have seen that the degree ranking was not outperformed by more elaborate ranking measures. However, the ranking did not seem to be very robust. Rankings on datasets reduced to one day proved to be slightly less efficient. How long data needs to be registered in order to provide a basis for robust degree-based immunization strategies and how much these strategies will vary depending on the aggregation time are therefore interesting questions, which we try to tackle here.

In order to understand which is the optimal aggregation time to arrive at a degree ranking which is sufficiently robust and efficient, it is important to understand how the ranking changes with longer aggregation times. The degree distribution changes with longer aggregation time towards larger average degrees and higher variance (see Fig. 2.5(a) and Fig. 2.5(b)) as new links are added to the network. We consider here the stability of the degree ranking of a network in which links instead of nodes are added, similarly to the discussion of leader nodes in growing networks [34]. The question is therefore if the ranking is preserved when new links are added or if it changes frequently and unpredictably. If the nodes have different inherent attractiveness, as expressed by their different degree in a static network, the ranking is expected to stay fairly stable.

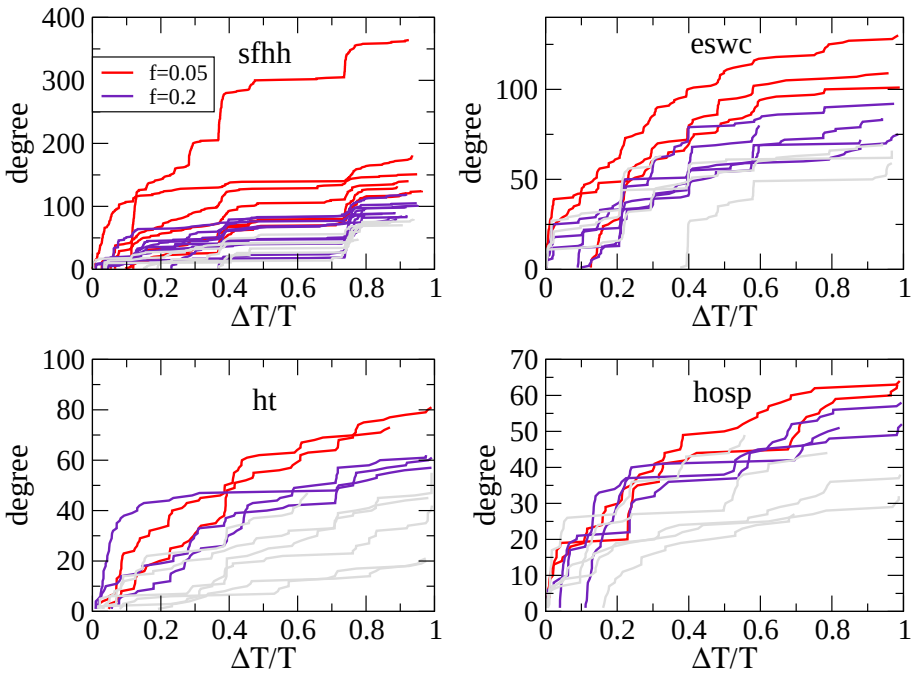


Figure 6.1: Degree of nodes on the network aggregated over ΔT vs. ΔT. The degrees of the first 5% of nodes in the degree ranking on the fully aggregated network are shown in red (only every third node is shown for clarity). The degree of the following nodes in the ranking (up to the first 20% of nodes) in the fully aggregated network are shown in violet (only every sixth node is shown). The evolution of the aggregated degree of a small number of other nodes is shown in grey for comparison.

We look at the development of the aggregated degree of the nodes as the network is aggregated over longer and longer time periods in Fig. 6.2 and 6.1. The dynamic networks used are "sfhh", "eswc", "ht" and "lyon2012" where in this section all nights and phases without activity were

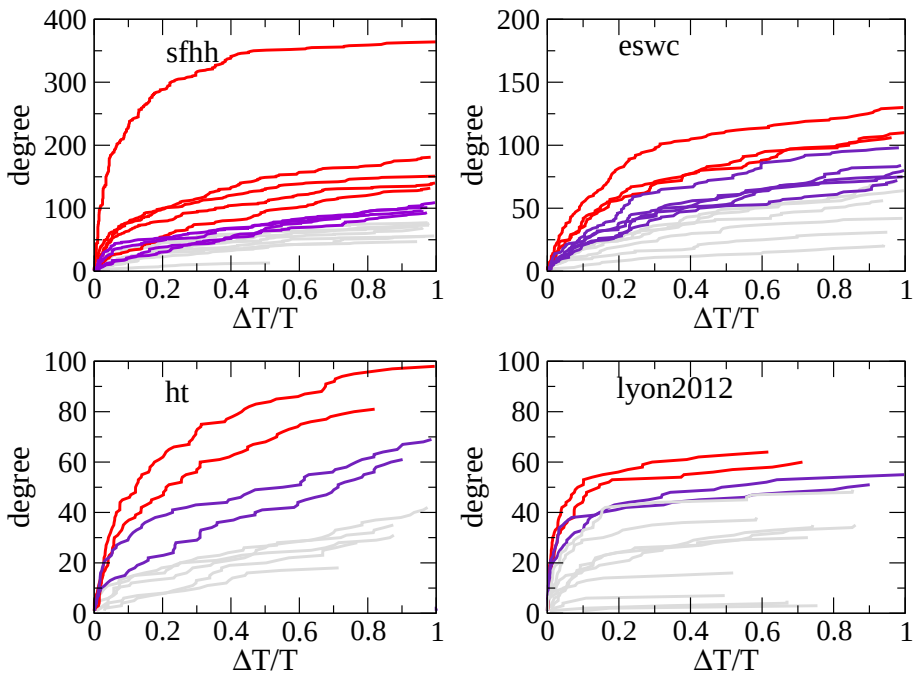


Figure 6.2: Degree of nodes on the network aggregated over ΔT vs. ΔT for a shuffled version of the dynamic contact data. See Fig. 6.1.

removed. In Fig. 6.2 each link has a specific probability of appearance. The dynamics of the network follows the Poisson distribution. In this case, the degree of the nodes grows in a regular way, which mostly preserves the ranking among different nodes.

The original data however has bursty dynamics. In Fig. 6.1 the development of the nodes' degree is plotted for the original network data. The growth of the degree with aggregation time shows irregularities and jumps, the ranking among different degrees is less stable. As the cutoff of the dynamic network data is arbitrary, the ranking most likely continues to change. In general, however, the ranking for the top nodes seems more stable than the ranking for lower degree nodes. The nodes which have the highest degree on the fully aggregated network (HET) already appear early on in the list of top nodes and there is little mixing across the whole range of the ranking. Possibly this stability can be a transient phenomenon as the datasets are rather short. For the hospital dataset, which is the longest of the four, the ranking seems to stabilize much later.

Two questions arise: What is the minimal aggregation time necessary to obtain a sensible ranking of influential nodes? And how stable is this ranking?

In Fig 6.3 we compare the ranking calculated on the network data aggregated up to time ΔT (ΔT-network) with the ranking on the fully aggregated network (T-network). Not all nodes are present over the entire length of the dataset. Nodes which have not appeared yet at time ΔT are randomly added at the bottom of the ranking. Instead of their ID, nodes in the ranking are referred to by their respective degree on the T-network. Nodes which have the same degree on the T-network, occupy the same position in the ranking. We then calculate the Kendall τ-b coefficient as a measure of comparison, as it takes account of ties. In a very short time, the Kendall-τ rises rapidly to a value above 0.5, as the fraction of nodes for which the ranking

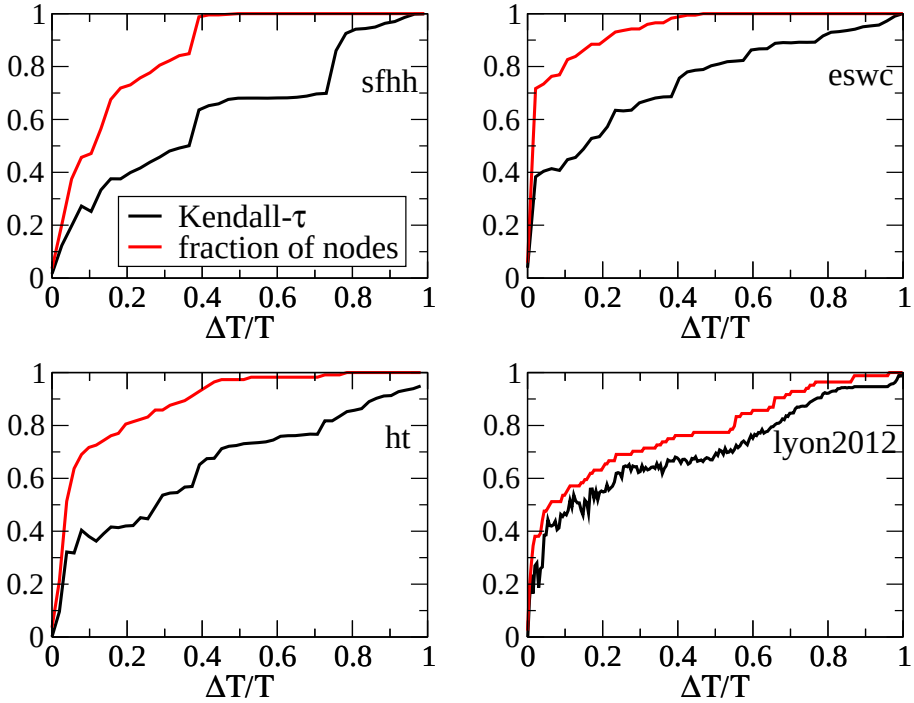


Figure 6.3: The Kendall τ-b correlation coefficient for the correlation between the ranking based on the fully aggregated network and the ranking based on the network aggregated up to ΔT.

is calculated increases. Afterwards, the ranking correlation coefficient only increases gradually. The most important factor seems to be the fraction of nodes which are ranked. As important nodes are supposed to be present more frequently than nodes which play a minor role in the spreading of epidemics, measuring contacts for only a short period could already be sufficient. Furthermore, as the density of nodes with similar degree is higher for low degree nodes, there are also more fluctuations in the bottom part of the ranking. The Kendall-τ ranking coefficient does not give different importance to the correct ranking of nodes in the top or the bottom part of the list. As low degree nodes do not play an influential role in the spread of epidemics, their exact ranking position is of minor importance.

In containment strategies, a fraction of the nodes of the network is removed. This fraction is the top fraction of the ranking. In order to find the optimal nodes for vaccination, it is therefore sufficient that they are ranked among the top nodes. Their exact position in the ranking is of no importance since for a given vaccination strategy a specific percentage of nodes is selected for vaccination.

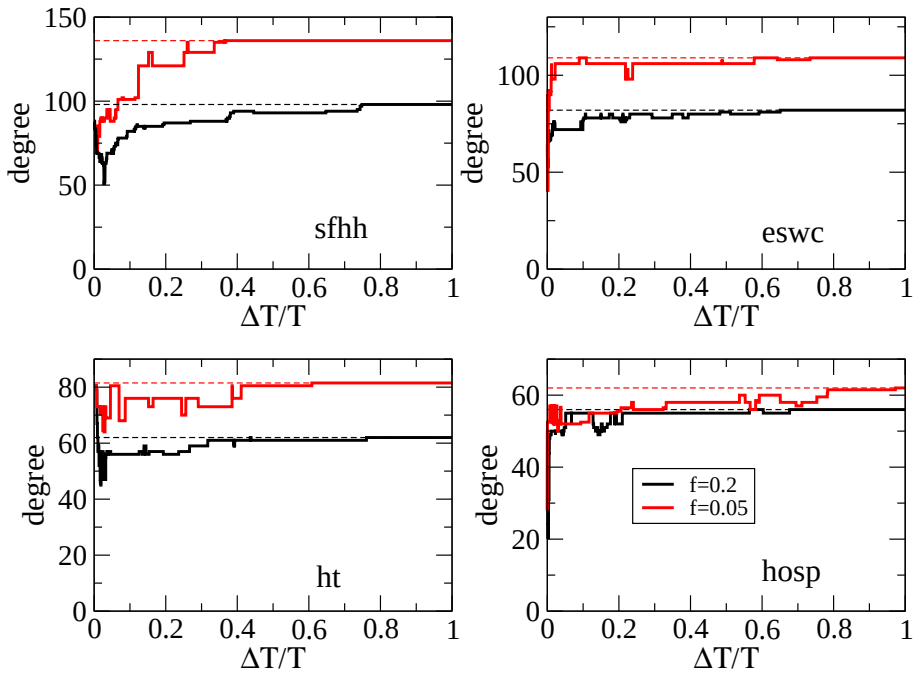


Figure 6.4: The median of the degrees of the top 5% (red) and the top 20% (black) nodes of the degree ranking based on the network aggregated up to ΔT, whereas the degrees taken to calculate the median are the respective degrees these nodes have *in the fully aggregated network*. The median degree is plotted vs. $\Delta T/T$. The dashed horizontal lines mark the final values: the median degree of the top 5% and of the top 20% of the nodes on the fully aggregated network.

In Fig. 6.4 we compare the top 5% and the top 20% of nodes of the degree ranking on the network aggregated over the complete length T of the data with the top nodes of the network aggregated over a time ΔT. For comparison we use the median degree of the nodes chosen at time ΔT, where the degree of each node is represented by its degree on the network aggregated over T. Already after a very short aggregation time there are many nodes among the chosen top nodes which play an important role on the entire dataset. The median degree of the top

nodes at time ΔT is rapidly similar to the median degree of the highest degree nodes calculated on the fully aggregated data. Aggregating for a longer time only slightly improves the choice of nodes. Whether the higher extra cost for a much longer collection of information is worth the improvement of the exact ranking of the nodes therefore depends on the stability of the ranking itself. If nodes had a fixed unchanging importance in the network, it could make sense to measure over a long enough time in order to find this ranking with high precision.

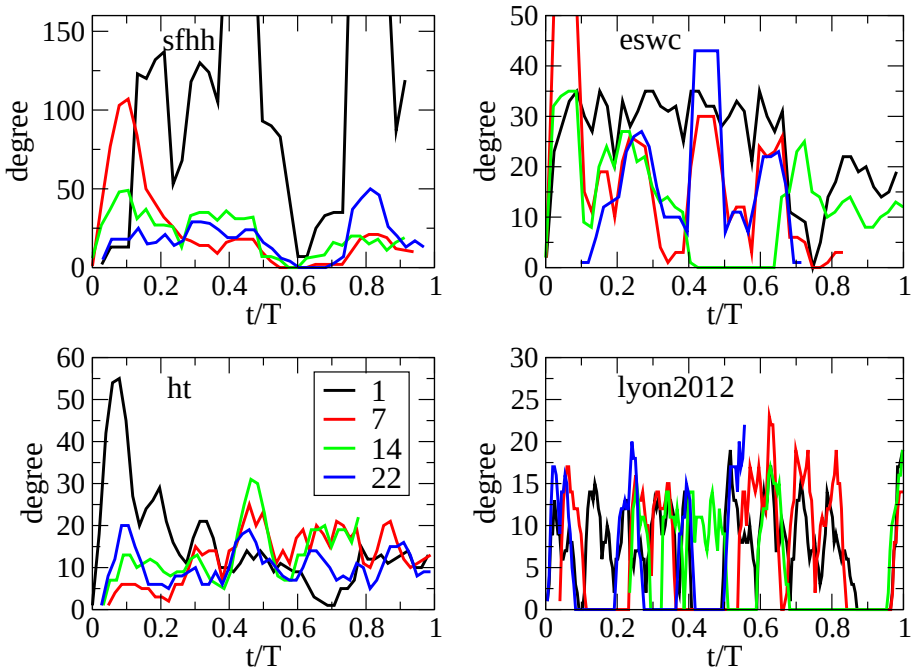
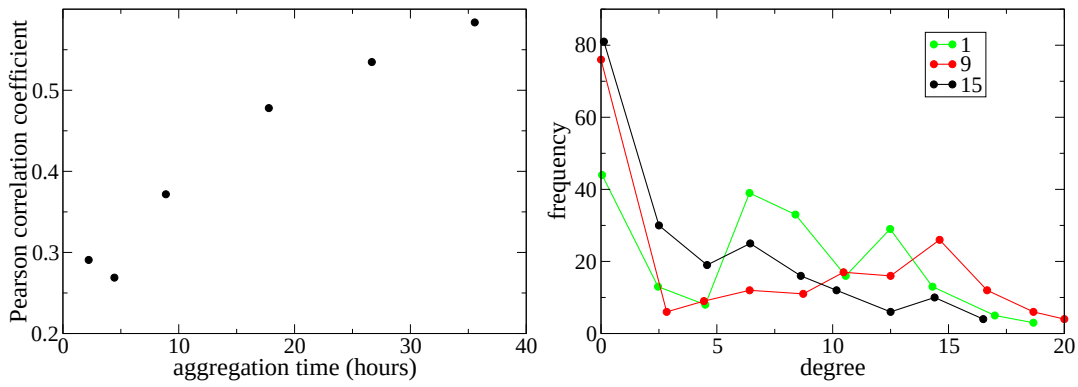


Figure 6.5: Degree of several nodes on the networks aggregated over temporal windows of 400 time steps vs t/T. The curves are colored according to the node's ranking in the network aggregated over $[0, T]$.

However, it was shown for an email contact network in [15] that this is not necessarily the case. The important nodes change frequently and do not occupy a crucial position in the dynamic network over a very long time period. In Fig. 6.5 we plot the aggregated degree of some of the top nodes depending on the starting time of aggregation. For a short aggregation time of 400 timesteps, the accumulated degree of nodes fluctuates strongly, especially in the hospital dataset. Even the importance of those nodes which occupy a top position in the fully aggregated network is localized in time. There are time periods in which these nodes do not appear at all. For longer aggregation times, the fluctuations are slightly attenuated. As the present datasets are rather short, it remains unclear, whether or not there are fluctuations on longer timescales.

In order to be able to predict if the top nodes which have been chosen by aggregating over a time ΔT will continue to play an important role later on, the correlation between the present and future degree of nodes needs to be known. We furthermore test the dependence of this correlation on the aggregation time. In order to test for different aggregation lengths, tests are done on the longest dataset, "lyon2012", where as before all phases without activity were taken off.

In Fig. 6.6(a) the Pearson correlation coefficient is plotted over the aggregation time. It



(a) (b)

Figure 6.6: (a) Pearson correlation coefficient between degrees of all nodes at time $t = 128000s$ and time $t = 320000s$, aggregated over different time length, vs aggregation time. (b) For nodes at position 1, 9 and 15 in the final degree ranking, the histogram of degrees at different points in the "lyon2012" dataset is shown. The degrees are calculated on networks aggregated over 400 timesteps (8000 s).

measures the correlation between the degree centrality at times $t_1 = 128000$ and $t_2 = 320000$. These points in time correspond to the last day of the first week and day four of the second week. Both show high activity. The degree was calculated on snapshots of the temporal network, aggregated over given time intervals with different aggregation times.

The correlation between the degree increases with longer aggregation time. We do not (yet) get a saturation effect as observed in [15], but a slight flattening is already visible. Correlation is in general much higher than in [15]. This could be related to the smaller size of the dataset.

For a selection of nodes we look at how much their aggregated degree varies. We aggregate the degree over 400 timesteps (∼2h.) Even though the degree distribution is very wide, we do not find a power law degree distribution as was found for phone calls by Braha [15]. Nodes which occupy a high position in the ranking on the complete dataset consistently show a small and wide peak at high degrees when considering only the time in which they are present. For nodes which occupy a lower position in the final ranking, this peak drifts towards zero. A second peak at degree zero is present for all nodes, as they are not constantly in contact with other nodes. The fact that the degree distribution does show a more consistent behaviour for high degree nodes, could be due to the limited length and size of the dataset but it could also be due to the fact that in the hospital dataset individuals occupy different roles, which distinguish them from each other.

To better visualize the development of the ranking for different roles, we plot in Fig. 6.7 again the degree as a function of aggregation time for the hospital datasets "lyon2012" and "obg", where the different roles are color coded. The behaviour of the curves is very distinct for the nursing staff and patients. While Nurses and Assistants have almost all been present in the first days, for some patients the first contacts happen much later. The degree rises rapidly for nurses and assistants, once they are present, while the degree for patients stays constantly low. It seems very unlikely that even after longer aggregation time, patients will adopt nurse-like behaviour or nurses patient-like behaviour. The average degree for the groups shows a robust ranking already for short aggregation times.

In conclusion, it can be said that for the datasets we consider here a preliminary ranking can

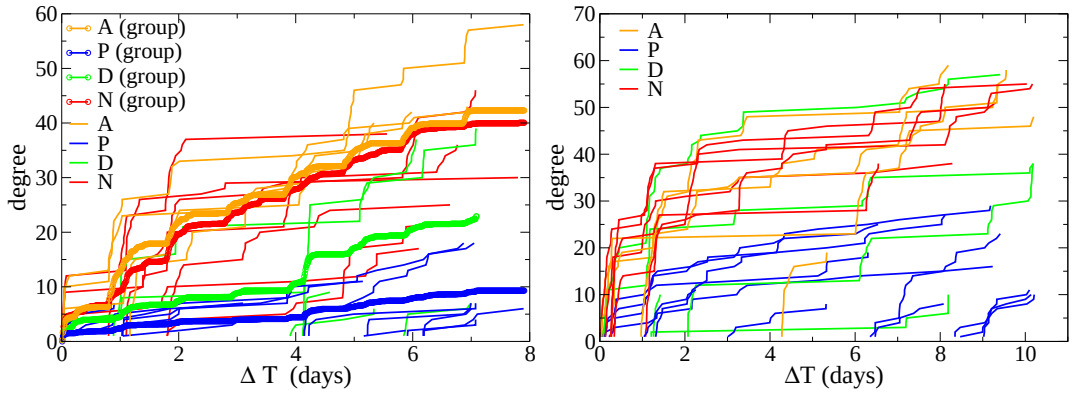


Figure 6.7: Left figure: Degree of nodes on the "obg" network aggregated over ΔT. The colored lines show the development of the degree of each individual belonging to a given class, the circles show the average degree for individuals of a given class. Right figure: Degree of nodes on the "lyon2012" network aggregated over ΔT. The different nodes are color coded according to the roles they occupy. The roles are: A- assistants, D- Doctors, N- Nurses, P- Patients. Only every third node of each role is shown.

be obtained after a very short time of data collection. This ranking already contains a major part of the most important nodes. Even though the ranking is not absolutely stable, the top nodes seem to reliably persist over large parts of the dataset, which might be, at least for the hospital dataset, related to the functions they perform. Furthermore, the ranking for groups proves to be very stable while the ranking for individuals fluctuates more prominently. The ranking is based on aggregated networks. The variations of the ranking therefore must come from variations of the underlying data. In the next section we compare two datasets taken at different times.

6.2 Data-based predictions of epidemic spread

Models are fed with data from a specific situation in the past. This situation-specific data will not repeat itself in the exact same way. The precision of the predictions therefore might depend on the precision of the data the model is fed, and on how much reality changes, that is, on how applicable the data is to the future situation. To make the data (in this case the face-to-face contact data) more general, knowledge about the process through which face-to-face contacts occur and the corresponding underlying distributions would be necessary. As of today, this process remains unknown. We use therefore the CMD data representation, described in the last chapter, as it is fairly general and keeps enough aspects of the data in order to allow reasonable predictions. Also, the size of the dataset can be easily modified in order to match the setting for which the prediction will be done.

We will ignore here the difference between the predictions based on the CMD data representation and an outcome of the epidemic in reality. We are only interested in the difference of the predictions due to data changes over time. To this end, we compare the predictions of the SIR-model on the CMD representations of two datasets which describe face-to-face contacts at the same location but at different times.

We test how well the model, informed by one dataset, can predict outcomes of epidemic spread in a different situation described by another dataset. The two datasets are the datasets "lyon2011" and "lyon2012" as described in chapter 2. To this end, we will first compare the two

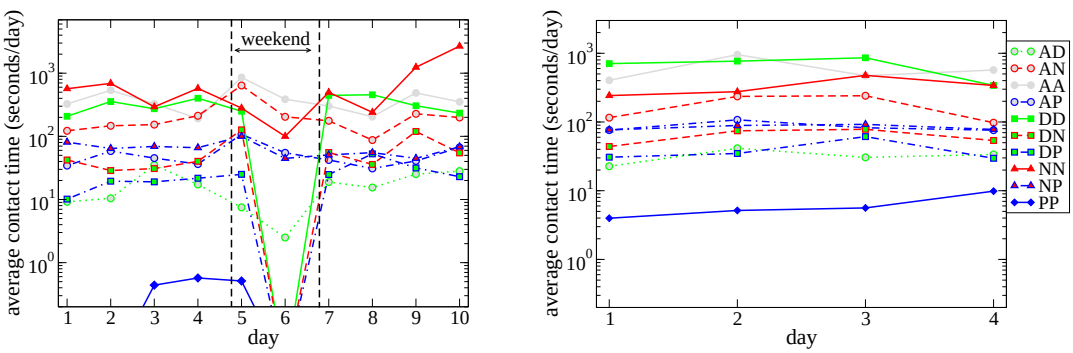
datasets in order to see how much data taken at the same location, but at different times, varies. Thus we can see how similar the data is, at a certain level of representation in 2 different years. Then we simulate epidemic processes on the data in order to see how important those differences are for the predictions based on the data.

6.2.1 Comparing datasets

Comparing the datasets on an individual basis is impractical since many participants are not present in both years. Furthermore, it cannot be expected that contact patterns of individuals are unchangeable. On the other hand, in order to make data-based predictions about future properties of the contact patterns, general features of the data should stay robust. We will therefore here only consider the CMD data representation, as it is a role-based representation with general rather than individual information. The participants of each dataset can be divided into 4 classes: Doctors, Nurses, Assistants and Patients.

The number of individuals which constitute one class is not equal in both datasets, nor are the exact composition and characteristics of the classes. Individuals can be assigned very specific tasks and roles. We group individuals with similar characteristics together. The class of nurses also includes student nurses and physiotherapists. The class of assistants includes nursing auxiliaries, student nursing auxiliaries, social workers and hospital service employees. The class of doctors is the most diverse, including among others an ergotherapist, a dietician, a psychologist and first and second year student doctors. Not all patients stay during the entire time of the data acquisition. Some leave and are replaced by other patients.

The CMD data representation uses the contact matrix and the matrix of distributions of the available dataset. If the general and essential features of the data are captured, and if they are constant, then the matrices should be the same for both datasets. When a network is constructed based on the CMD data representation, group sizes can be chosen corresponding to the group sizes of the dataset for which the predictions are done.



(a) lyon2012 (b) lyon2011

Figure 6.8: The average contact time in seconds/day, calculated on daily contact matrices for the "lyon2012" and the "lyon2011" dataset, vs the day on which it is based. Lines are guides to the eye. Contact matrix entries for contacts among the same group are marked by a continuous line. Symbols have one outline and line color which represents one of the two interaction partners and one filling color and symbol, which represents the other interaction partner. Nurses - red (triangle), Doctors - green (square), Patients - blue (diamond), Assistants -grey (circle)

In Fig. 6.8 we show the average contact time for each day between two people of one of

class	2011	2012	
		week 1	week 2
A	15	12	12
D	11	16	15
N	16	11	12
P	29	21	28
all	71	60	67

Table 6.1: Number of individuals in each class during the 4-day period

class	2011				2012									
day	1	2	3	4	1	2	3	4	5	6	7	8	9	10
A	12	11	12	12	10	8	9	11	6	8	8	11	9	8
D	10	10	9	10	10	12	9	10	4	1	11	12	9	11
N	12	9	9	11	9	8	9	9	3	6	8	9	9	9
P	19	20	21	17	14	17	14	15	13	14	18	18	17	16
all	53	50	51	50	43	45	41	45	26	29	45	50	44	44

Table 6.2: Number of individuals per day in each class

the classes, measured in seconds per day for the "lyon2012" hospital data (6.8(a)) and for the "lyon2011" hospital data (6.8(b)). Fluctuations are clearly visible between different days, especially for the weekend when contact times change dramatically. Doctors are not present on Sunday, and assistants have higher average contact times on Saturday. However, during the weekdays the ranking between the classes is fairly stable.

For a better comparison between the two datasets, the datasets will be cut into parts of equal duration, which only consist of weekdays, excluding the weekend in the "lyon2012" dataset. For the "lyon2012" dataset, we consider only the first four days (week 1: "lyon2012_w1") from Tuesday to Friday and the last four days (week 2: "lyon2012_w2") from Monday to Thursday. The "lyon2011" dataset spans over the duration of four consecutive days from Monday afternoon until Friday noon.

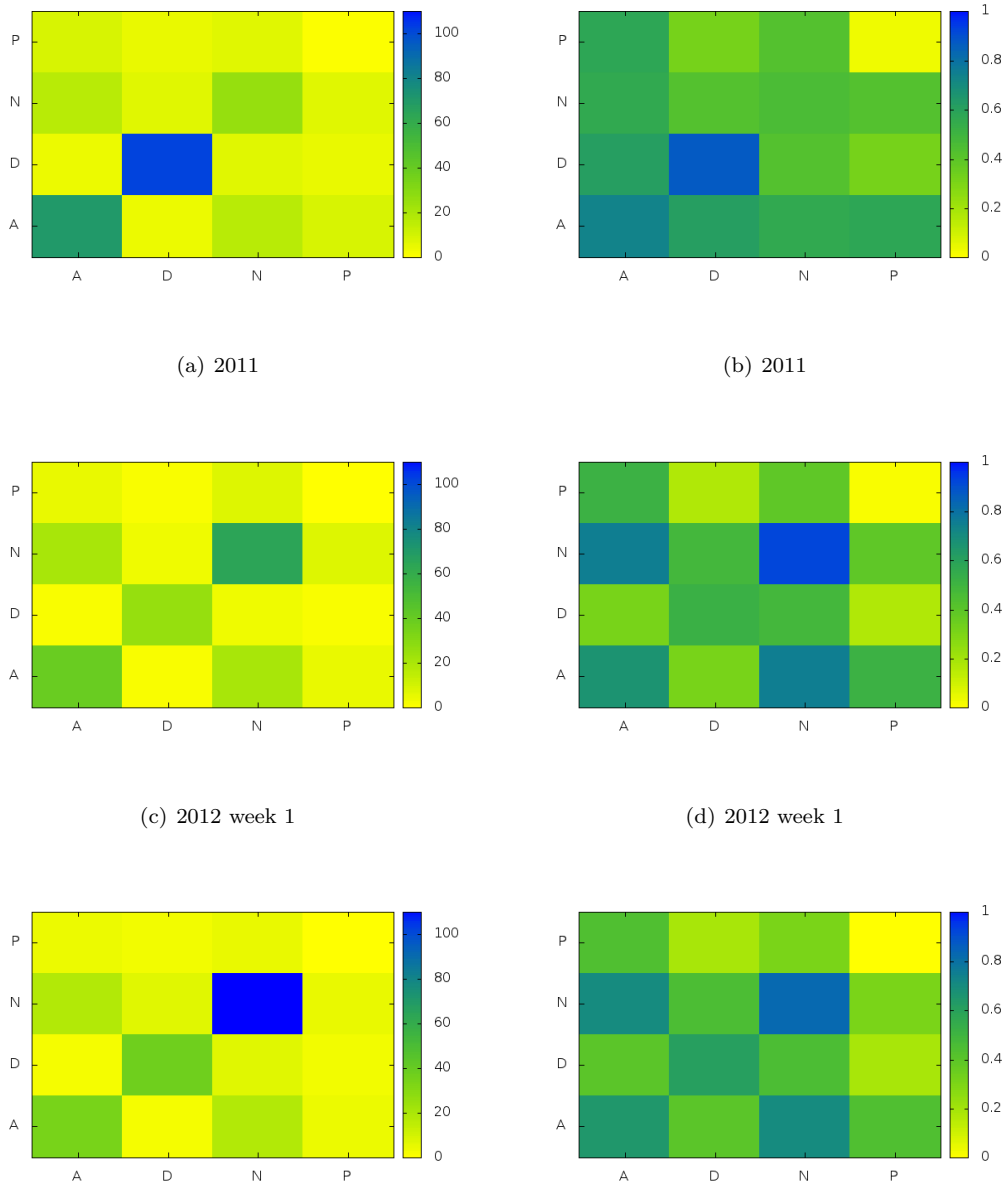
The class sizes in the three four-day networks are similar (see Tab. 6.1). Class sizes fluctuate strongly on a daily basis, especially over the weekend, where they are much reduced (see Tab. 6.2).

The fact that not all individuals of each class are present every day leads to lower daily average contact times between classes for the 4-day networks as compared to the daily average contact time on the daily networks, where absent individuals are not used to compute the average.

The general features of the contact matrix are comparable to the contact matrix of the "obg" dataset from the children's hospital in Rome (see Isella et al. [40] and Tab. A.1). Contacts between patients are very low, contact times and density between assistants and nurses are quite high.

The contact matrices and the link density in Fig. 6.9 show that the two weeks in the "lyon2012" dataset are very similar. The only difference is a strong rise in the Nurse-Nurse contact time, as already observed towards the end of week 2 in Fig. 6.8. The "lyon2011" dataset, however, shows very different properties. Here the contacts among doctors are extremely high. This is mainly due to two doctors talking about two hours every day. At this time, student doctors, who needed some tutoring, were in the hospital. The link density is also higher (Fig. 6.9(b), suggesting more interaction among doctors in general.

Overall, the three datasets show slightly different properties, especially the high activity



(a) 2011

(b) 2011

(c) 2012 week 1

(d) 2012 week 1

(e) 2012 week 2

(f) 2012 week 2

Figure 6.9: Left column: average contact time between individuals of different classes (A: Assistants, D: Doctors, N: Nurses, P: Patients) in seconds per day. Right column: density of links between the different classes.

for doctors in the "lyon2011" set in comparison to the "lyon2012" sets. On the other hand, the average contact times between groups seem to be very similar for the majority of groups. Whether these differences have an effect on the outcome of the epidemic will be tested in the next section.

6.2.2 Effect of data variability on epidemic predictions

If the predictions of datasets are comparable, then any of the given CMD data representations can be used in order to predict the outcome of epidemics on any of the given datasets. In Fig. 6.10(a), 6.10(c) and 6.10(e) we compare the outcome of the epidemic spread using the CMD data representation fed by one of the three datasets as a prediction for the outcome on the other datasets. The parameters used for the simulation are: $\beta = 100\mu$, $\gamma = 1/2\text{day}^{-1}$, $\mu = 1\text{day}^{-1}$ The outcomes of the curves marked as "2011 $\rightarrow$ 2011", "2012_w1 $\rightarrow$ 2012_w1" and "2012_w2 $\rightarrow$ 2012_w2" are the standards which we try to obtain by using the CMD representation of the respective other datasets. They are already quite different from each other. Many factors can play a role for this difference. The number of participants is not identical, the number of individuals per group differs and the contact matrices do not contain the same average weight for entries between groups. When using, for example, the CMD representation of the "lyon2011" dataset to simulate the epidemics for the "lyon2012_w1" dataset, we create a network with the number of participants and individuals per group taken from the "lyon2012_w1" data and the weight distributions and corresponding average weight of the CMD representation of the "lyon2011" data. The outcome of this simulation is marked as "2011 $\rightarrow$ 2012_w1". Thus eventual differences due to a different number of participants are eliminated.

The influence of the group sizes can be seen by comparing the outcome of simulations which use the same data to create the contact matrix of distributions, but different group sizes. The effect is generally small. It is most visible by comparing the simulation in the case where the CMD matrix is based on the "lyon2012_w1" data set, while group sizes are taken either from the "lyon2011" or the "lyon2012_w1" dataset.

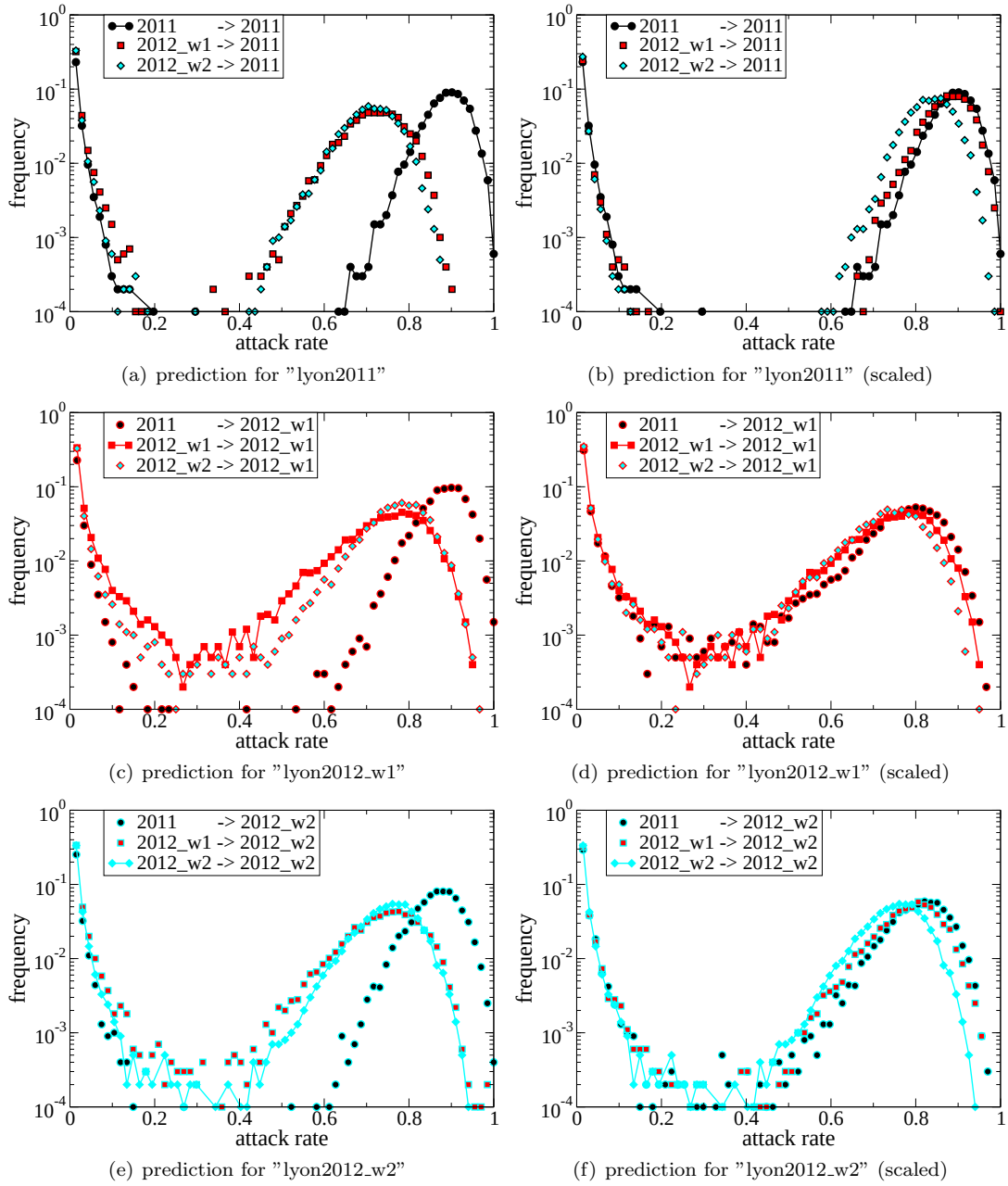
The simulation on the first week of "lyon2012" gives a very good prediction for the outcome on the second week of "lyon2012". Using the second week of "lyon2012" as a prediction for the outcome on the first week of "lyon2012" still works well. However, any prediction of the final size of the epidemic for the "lyon2011" dataset using one of the two weekly sets of the "lyon2012" set is poor, the final size is largely underestimated. In the opposite case, the spread simulated for an epidemic in "lyon2012" using data of "lyon2011" is overestimated.

The strong difference in the average final size of the epidemic between simulations using the CMD representations of the 2011 and the 2012 datasets is mainly due to the difference in average weight of the contact matrices.

The average weight of the 2011 dataset is much higher than that of the 2012 datasets. The average weight was: "lyon2011": 62 s/day, "lyon2012" week 1: 46 s/day, "lyon2012" week 2: 51 s/day.

If in addition to adjusting the group sizes, we also adjust the average weight of the complete network, the average outcome is much better (Fig. 6.10(b), 6.10(d) and 6.10(f)). However the estimate for the importance of classes is still different (see Fig. 6.11).

This difference in the relative importance of classes is directly attributable to the structure of the different contact matrices. When comparing the daily entries of the contact matrices in Fig. 6.8(a), then the difference between the contact matrix entries in the "lyon2012" and the "lyon2011" dataset, especially for the average contact time among nurses and the average contact time among doctors, seems outside of the normal day-to-day fluctuations. However, as the two datasets are rather short, it is not possible to tell what distribution the day-to-day fluctuations



(a) prediction for "lyon2011"

(b) prediction for "lyon2011" (scaled)

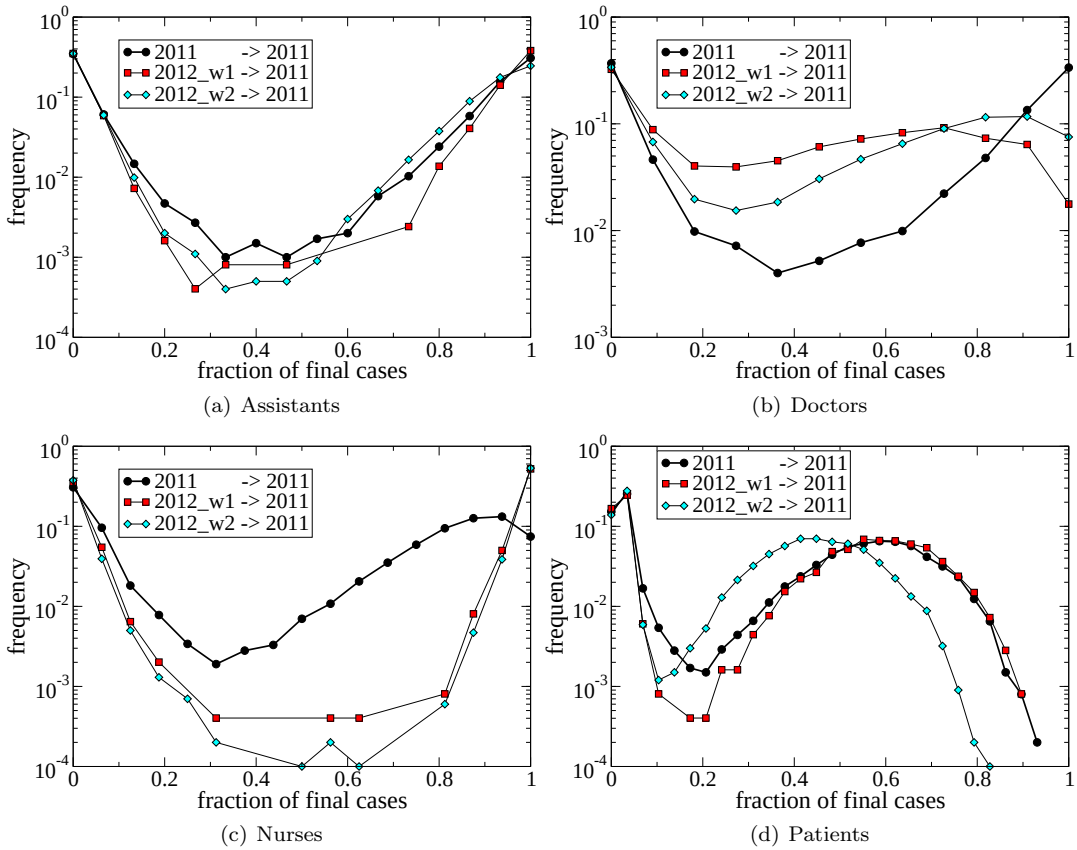
(c) prediction for "lyon2012_w1"

(d) prediction for "lyon2012_w1" (scaled)

(e) prediction for "lyon2012_w2"

(f) prediction for "lyon2012_w2" (scaled)

Figure 6.10: The final size of the epidemic, simulated on the CMD data representation with parameters $\beta = 100 * \mu$, $\gamma = 1/2\text{day}^{-1}$, $\mu = 1\text{day}^{-1}$. The legend "2011" $-$ > "2012_{w1}" means, that the simulation was done using the contact matrix with distributions calculated on the "lyon2011" dataset with group sizes adjusted in order to serve as prediction for the first week (w1) of the "lyon2012" dataset. Similarly for all other combinations of datasets the form of the symbol and color is characteristic for the dataset which is used to create the contact matrix of distributions, while the symbol outline stands for the dataset for which the simulation is done. If the dataset for which the simulation is done is the same as the dataset on which the contact matrix is based, the symbols in the plot are connected by a line. The histogram is built from 10000 simulations. Every 100 runs, the network is rebuilt from the CMD representation of the data. Right column: the weights of the resulting network were additionally scaled in order to show the same average weight as the network for which the prediction is done.



(a) Assistants (b) Doctors

(c) Nurses (d) Patients

Figure 6.11: The final size of the epidemic, simulated on the CMD data representation with parameters $\beta = 60 * \mu$, $\gamma = 1/2\text{day}^{-1}$, $\mu = 1\text{day}^{-1}$. The fraction of final cases in each group Assistants, Doctors, Nurses and Patients are shown. Simulations are done with the group sizes of the "lyon2011" dataset and the weights are rescaled to have the same total average weight as the "lyon2011" data. The weight distributions are calculated using the "lyon2011" data, the "lyon2012_w1" and the "lyon2012_w2" data.

follow. Therefore, at this point we cannot decide whether the change in average contact times between the datasets is within the normal day-to-day fluctuations or if it constitutes an abnormal exception, a change in hospital procedures. The distribution of weights among and between groups is rather large and modeled here by negative binomial distributions. The average weight changes from day to day or week to week. If the process which generates the daily contact weights had every day the same underlying negative binomial probability distribution, the distribution of daily average weights or contact times between two classes could fluctuate broadly. Very long datasets or very large groups would be needed in order for the average contact time to stabilize. However, if the underlying distribution of contact times or weights does not have an expectation value or if the distribution is very broad, then increasing the sample size within given possibilities would not stabilize the result. Furthermore, with the increase of group sizes or the aggregation time of the data simulations on the contact matrix representation become less similar to simulations on the original data. This is a trade-off which needs to be looked at more closely in order to evaluate the faithfulness of predictions using the CMD data representation. With longer data collections, however, a distribution of daily contact matrices could be approximated. Knowing the variance of this distribution can lead to a better understanding of the precision of predictions based on available data. For a given daily contact matrix it could then be decided if it is an outlier of the distribution or not.

6.3 Conclusion

In order to estimate the variability of epidemic predictions, not only the stochasticity of the epidemic model plays a role, also the variability of the underlying data, on which predictions are based, needs to be taken into account. Here and in the last section we have seen that the degree ranking of face-to-face contact data varies strongly for short aggregation times on an individual basis. Degree ranking on a group basis is more robust. However, even when aggregating the datasets over longer times, here over 4 days, and only considering group properties, the outcome of the epidemic is significantly different for datasets which lie a long time apart. This has direct consequences on the possibility to predict future epidemics based on present data or epidemics on different settings. Even when the right number of participants is known, the frequency of their interactions can change, leading to a wrongly estimated importance for groups. It is therefore important to have an estimate for the variability of the contact patterns in addition to the variability of the epidemic spread.

Chapter 7

Distances

In temporal networks, the spread of information or diseases has to follow the time ordered events. Unlike the spread on static networks, temporal constraints can hinder the direct transmission between two adjacent nodes. If this is the case, then the flow of information or pathogens between two neighboring nodes is only possible via third parties, if at all. This can lead to much longer transmission paths than in static networks. As information can accumulate errors at every retransmission, the number of steps it has taken before arrival can severely influence its accuracy and thus its trustworthiness. Finally, some processes on networks, like the propagation of a certain behaviour, only permeate to finite depths. In all these cases, the path length of the process is a valuable information.

The distance between two nodes on the static network is a lower limit for the path length of any process on the network. It gives a measure for the connectedness between two nodes. In dynamic networks the distance between two nodes is not necessarily identical to the length of the fastest path, also called temporal path. While the first minimizes the number of steps between two nodes, the second minimizes the time information or pathogens need to travel between two nodes. Just as the distance on the static network indicates how well two nodes are connected, so does the temporal path length on the dynamic network for a specific time. The temporal path can be characterized by the number of intermediate nodes which the temporal path traverses (temporal path length) and the time it takes to travel along the temporal path. This time can be very diverse and is only loosely correlated with the distance on the static network [73]. In Sec. 7.1, we will show that this also holds for the temporal path length.

Therefore, the distance on the static network cannot sufficiently inform about the number of steps information travels between two nodes on the dynamic network. In order to know how many individuals will be reached within a limited amount of steps, more knowledge about the distribution of temporal path lengths is needed.

In Sec. 7.2, we will try to compare the distribution of temporal path lengths with the distribution of infection-path lengths on static networks, since often only an aggregated static version of a temporal networks is available, but also in order to see the influence of temporal properties on the distribution of temporal paths. The infection-path length between two nodes is the number of steps which an SI process with $\beta < 1$ starting from one node takes to reach the other node. Infection paths, like temporal paths, are self-avoiding. Every node can only be visited once. The infection-path length on static networks is closely related to the temporal path length on networks with Poissonian dynamics. By comparing the path length distribution of temporal networks with Poissonian dynamics with the distribution of the original network, the influence of temporal properties, like time resolution, burstiness or correlation of events on the distribution of

temporal path lengths becomes apparent (see Sec. 7.3). Furthermore, the infection-path lengths on fully connected networks can be well approximated by analytic calculations, thus giving a good comparison to the simulated infection-path length distributions (see Sec. 7.2.1).

On a static network, the shortest path is the most likely path connecting source and target. However, depending on the network structure, many longer transmission pathways are possible as well. In fact, it is much more likely that the spreading process takes a considerably longer path [71]. So, even though every single longer path is less likely taken than the shortest path, not taking the shortest path can be more likely than taking it, depending on the number of other possibilities. The probability of taking longer paths therefore depends on the number of possible longer paths (and their respective probabilities). The number of possible transmission paths between two nodes increases with the amount of loops in the network. On networks with tree-like structure, the shortest path is the only available path. Therefore, the path length of epidemic processes will not differ from the distance on the static network if the network is a tree. On all other networks, anything between the minimum and the maximum path can be taken by the epidemic process when propagating from one node to another.

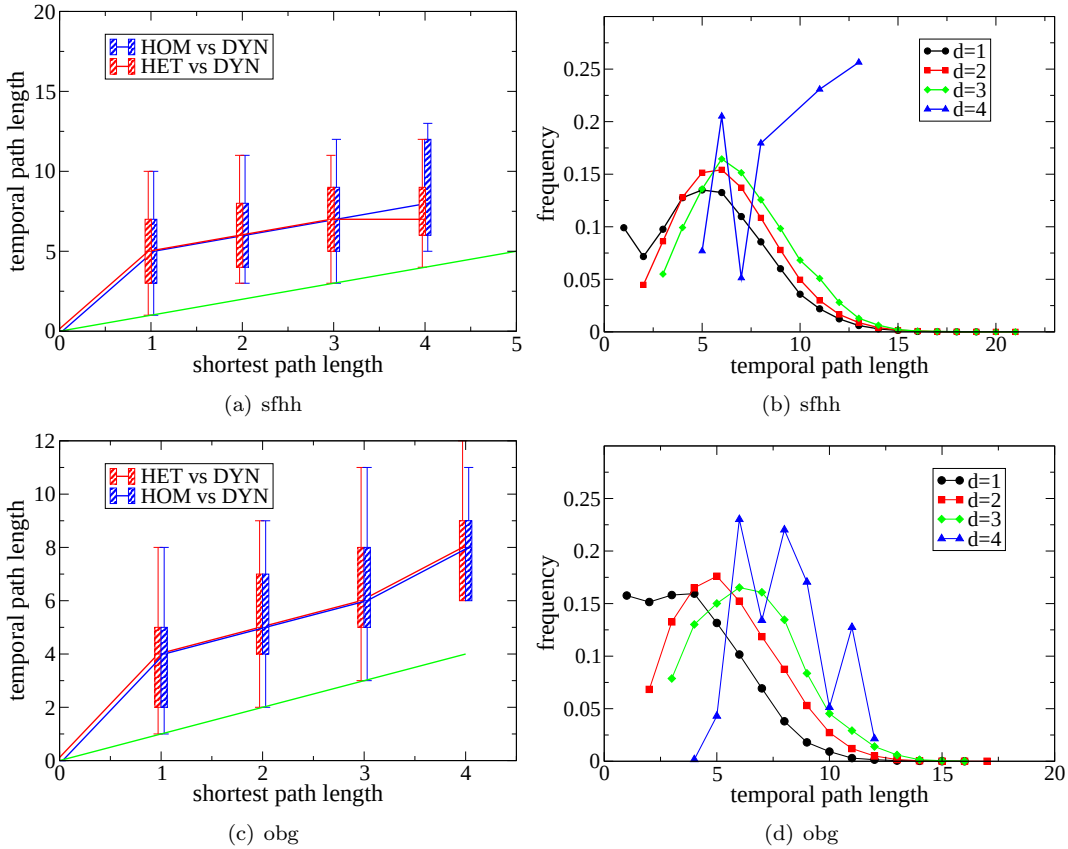
The relation between the infection-path length and the shortest path can inform about the structure of the static network, and the structure of the network can give insights about the distribution of the infection-path lengths. In the following, we will consider the influence of link density (in Sec. 7.2.2) and of the weight distribution (in Sec. 7.2.3) on the distribution of the infection-path lengths. The influence of network topology remains an open question.

7.1 Static distance vs. dynamic distance

In social networks, most individuals are connected via only a few intermediate friends. This property, which characterizes small world networks, has as a consequence that information could in principle travel quickly between any two nodes, following a very short path. In reality, individuals are not constantly in contact with each other. The fastest connection between two nodes, the temporal path, does not necessarily follow the shortest path. As a measure of how close two nodes are on a dynamic network at a specific time period and how fast information can spread between them, the temporal path [73] is a much more adequate measure than the shortest path. It provides a lower limit for the time something needs to spread between two nodes at a given time period.

We calculate the temporal path between any two nodes using an SI process with $\beta = 1$ on the dynamic network. Since the temporal path also depends on the starting time, for each node pair i, j we calculate the temporal path for 20 random starting points in the network. As starting time we chose the time of first occurrence of the seed i after the random starting point. Since the likelihood of reaching other nodes in the network decreases, as the start of the temporal paths is close to the end of the finite data set [73], we repeat the data once. For every seed node, the SI process will construct a spanning tree of temporal paths. The temporal path between node i and j is different from the temporal path between node j and i. It can be considerably longer or shorter. Temporal paths do not even necessarily exist in both directions.

In Fig. 7.1 we compare the temporal path length to the static path length on networks of face-to-face contacts from a conference ("sfhh") and a hospital ("obg"). The distances on the static network are very different from those on the dynamic network. The number of intermediate nodes on the temporal path is much higher than the shortest path length on the static network. The latter is only a lower bound. In dynamic networks, the duration of the path between two nodes and the path length are unlikely to both be minimum at the same time. If information is to be transmitted directly between two nodes, or via the fewest intermediate nodes, it can



(a) sfhh (b) sfhh

(c) obg (d) obg

Figure 7.1: Left: static distance vs temporal path lengths for the "sfhh" and the "obg" data sets. The distance on the static network was either taken on the HOM network or on the HET network. The temporal distance on the DYN network was measured starting at the first occurrence of each node. For better visibility, the 'HOM vs DYN' was shifted by 0.03 to the right, and 'HET vs DYN' was shifted by 0.03 to the left. The green line marks the identity and thus is a lower limit for the temporal distance. Right: distribution of temporal path lengths between nodes which have a given static distance.

take considerably longer than if it travels the fastest path, depending on the starting time of the process. If, on the other hand, the time to reach one node from another is minimized, as in temporal paths, the number of intermediate steps cannot be minimized at the same time and will therefore often be higher than the shortest path length. The impression of closeness between two nodes on the static network can thus be misleading. In many cases, information needs to travel a long way in order to arrive in the fastest possible way. However, the shortest path length is correlated to the temporal path length. The correlation became stronger by taking the average over several temporal paths between two nodes, starting at different points on the dynamic data. For any two nodes which are the same distance apart on the static network, the temporal path length on the dynamic network is distributed as shown in the right column of Fig. 7.1. Nodes which are direct neighbors are still likely to connect directly in the dynamic network, whereas information between nodes which are only separated by two steps on the static network is most likely to travel around five steps on the dynamic networks used here.

7.2 Temporal path lengths and infection-path lengths

Even in a densely connected network, infection can travel long distances before reaching individuals who are only a few steps away on the static network [71]. In the case of mutating pathogens or epidemics which only spread up to a finite depth, it is therefore of interest to know how many nodes can be reached up to a given path length. In order to know how many steps on average a process will take before arriving at a random node on the network, we will look at the distribution of temporal path lengths on dynamic networks.

The role the dynamics of the network plays for the temporal path length distribution can be assessed by comparing it to the distribution of distances in a network with shuffled time events. We will use a dynamic network where the links are active at random times with the same average probability as on the original dynamic network. This is the dHET network mentioned before (see Sec. 3.5). The temporal path on the dHET network can be simulated in the same way as the infection path on the static HET network (see Sec. 2.6). The infection path on a static network is the path of an SI process with $\beta < 1$. The distribution of the infection-path length and the temporal path length would be identical if simulations of the infection-path length were discrete in time and the propagation probability was given by the link weights $w < 1$. The temporal paths on the dHET network only differ from infection paths on the corresponding static HET network due to the discreteness in time of the temporal network. We simplify the underlying network structure further, looking at infection-path lengths on random graphs and fully connected networks. The most basic approximation, using the differential equations for an SI process, turns out to be a reasonable first approximation for the distribution of temporal path lengths in networks with low activity.

7.2.1 Discrete vs continuous

To understand the basic properties of the distribution of infection-path lengths, we look at the development of the SI process over time for nodes at different infection path steps from the seed. The simplest implementation, which can be solved analytically, is the homogeneous mixing case. Every node is connected with every other node so that the diameter of the network is exactly 1. There are no topological constraints for the spread of an epidemic, every path is possible. This leads to a maximal number of possible paths of any path length d.

In order to know at how many transmission steps from the source an infected individual is, the compartment of infectious can be divided into sub-compartments of infectious at a certain

transmission distance from the source. The differential equation for the percentage of individuals, which were infected after d steps,

$$\begin{aligned}\frac{\mathrm{d}i_d}{\mathrm{d}t} &= \beta n i_{d-1} s \\ &= \beta n i_{d-1}\left(1+e^{\beta n t}(\frac{1}{s_0}-1)\right)^{-1}\end{aligned}$$

can be solved recursively, using the solution of the SI model (Eq. 1.5). Initially only one seed is infected, so that $i_0 = 1/n$ and $s_0 = (n-1)/n$. All nodes which are directly infected by the seed are at distance $d = 1$. Infected at distance $d = 2$ have been infected by any one of the nodes at distance $d = 1$.

$$\begin{aligned}i_1(t) &= \beta n i_0 \int_0^t s(t')\,\mathrm{d}t' \\ &= \beta\frac{n}{n}S(t) \\ &= \beta\frac{\beta n t - \ln\left(1+(n-1)^{-1}\exp(\beta n t)\right)}{\beta n} \\ i_d(t) &= \frac{(n\beta)^d S(t)^d}{d!n}\end{aligned}$$

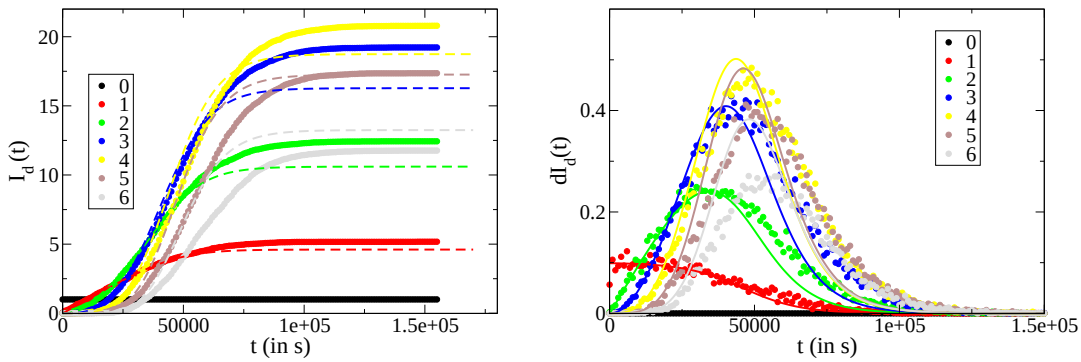
The solution to the differential equations is plotted in Fig. 7.2 for $\beta = 10^{-6}$ and $n = 100$. The number of infectious at distance one rises first. At some point, there are so many more infected nodes at a distance greater than one that it becomes more likely for nodes to be infected by nodes of a distance greater than one than directly by the source. The more nodes are infected at d steps away from the source, the more the number of nodes at $d+1$ steps away from the source grows. As the epidemic spreads, each compartment has its maximal number of newly infected individuals at a later time than the compartment from which it got infected. Therefore the average infection distance will increase with the number of nodes in the network. Ultimately all nodes are infected.

Together with the analytic solution of the differential equations, the simulation of a SI process on a fully connected network of 100 nodes is plotted in Fig. 7.2. The SI process on the fully connected network is stochastic and discrete. No partially infected nodes exist. Nodes will either get infected completely or not at all. At any time step, neighbors can get infected with probability $\beta = 0.00001$. The results are averaged over 1200 runs. The initial growth of the discrete process on the network is therefore slower than the growth of the analytic solution. The difference propagates through the different compartments so that with increasing spreading distance from the source the number of newly infected nodes peaks much earlier for the continuous process than for the discrete process, ultimately leading to a higher average distance from the source for nodes infected via the continuous process.

The final number of infected $I_d(t) = n i_d(t)$ at distance d can be found for the continuous solution by looking at the limit of $t \to \infty$:

$$\lim_{t\to\infty} I_d(t) = \frac{\ln(n-1)^d}{d!} \tag{7.1}$$

For the discrete solution, the probability to find nodes at a distance d can be calculated looking at all possible spreading trees. The probabilities to take the maximum path length, the Hamiltonian path, is $n/n!$ (if at every time step only one node gets infected). This is much lower than the probability for maximum path length in the continuous model if $n > 5$: $\frac{\ln(n-1)^{n-1}}{n!}$.



(a) n= 100, $\beta = 0.000001$ (b) n= 100, $\beta = 0.000001$

Figure 7.2: Left column: development over time of the number of nodes infected after d steps. Only curves for temporal path lengths up to $d = 6$ are displayed. The lines show the solution of the SI model, the dots the simulation on a fully connected network with $\beta = 10^{-6}$. Right column: development over time of the number of newly infected nodes at a temporal path length of d steps from the source. Only curves for temporal path length up to $d = 6$ are displayed. Lines show the SI model solution, dots the simulation on a fully connected network with $\beta = 10^{-6}$.

The average distance the epidemic travels is:

$$\begin{aligned}
\langle d \rangle &= \frac{1}{n} \sum_{d=0}^{\infty} d \frac{\ln(n-1)^d}{d!} \\
&= \frac{1}{n} \ln(n-1) \sum_{d=1}^{\infty} \frac{\ln(n-1)^{d-1}}{(d-1)!} \\
&= \frac{n-1}{n} \ln(n-1) \\
&\sim \ln(n)
\end{aligned}$$

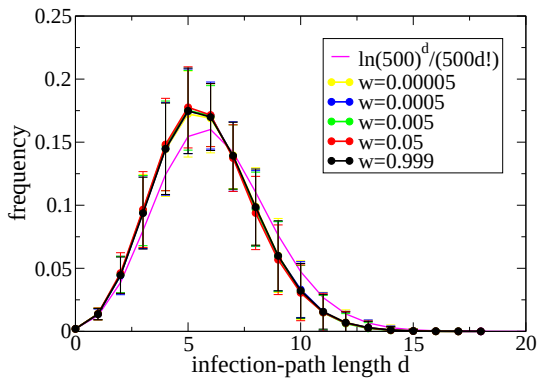
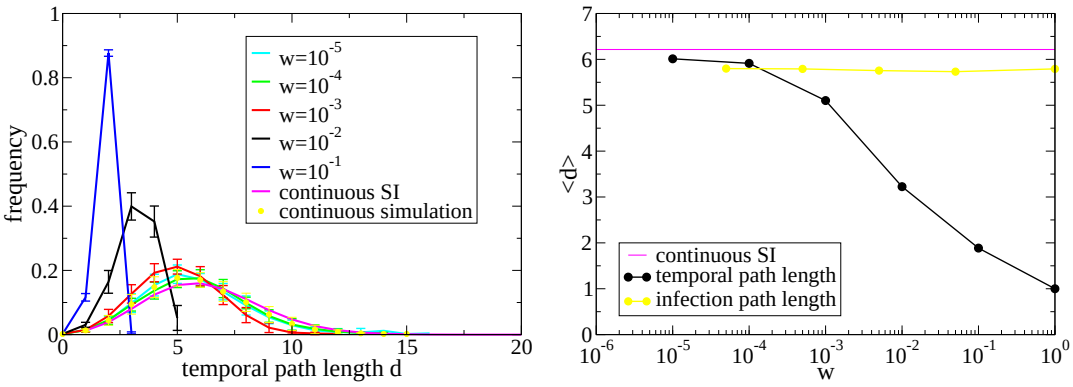


Figure 7.3: Simulation of the infection-path length distribution on a fully connected static network of 500 nodes for various β in continuous time. The solution of the SI model is shown as comparison.

In Fig. 7.3 the distribution of the infection-path length is plotted for the continuous and the discrete stochastic SI process. The distance distribution of the analytic solution is independent of β. Simulating the spreading process in continuous time but with discrete nodes for different β does not show any dependence on β either. As static networks do not have intrinsic timescales, any timescale of the process can be rescaled. Therefore only time related effects, like the change of the epidemic duration, can come from changing β.

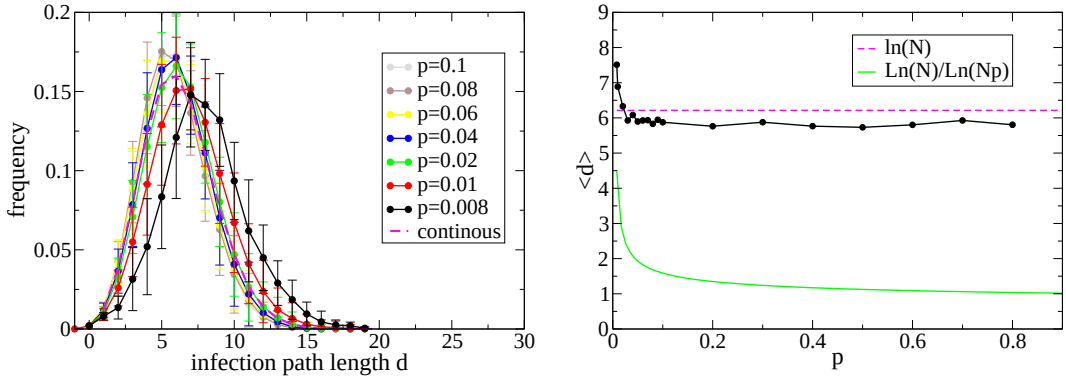


(a) (b)

Figure 7.4: (a) Simulation of the temporal path length distribution on a fully connected network of 500 nodes for various link activation probabilities w in discrete time. The simulation of the infection-path length in continuous time and the infection-path length of the SI model are shown as a reference. (b) The average temporal distance depending on w. Simulations in continuous time and the solution of the SI model are shown as a reference.

In contrast to static networks, dynamic networks have a timescale. Changing the timescale of the process on the network can change the outcome of the process. For temporal paths, the process on the network is instantaneous with $\beta = 1$. If we evaluate the propagation of the epidemic on the static network at discrete time steps, it is equivalent to an SI process with $\beta = 1$ on a dynamic network where the dynamics are regulated by a probability w for the links to be either active or inactive. In Fig. 7.4 we gradually change this probability w from 1 to 10^{-5}. If the probability for links to be active is 1, then the distribution of temporal path lengths is identical to the distribution of shortest path lengths on the static network, as we effectively have a static network. By lowering w we tune the network from one where the distribution of temporal paths is identical to the distance distribution on static networks to one where the distribution of temporal paths mimics the distribution of infection-path lengths on the static network. The average temporal path length increases as the instantaneous average degree wN decreases. The instantaneous degree of a node regulates how many of its neighbors can be infected at the same time. A high number of simultaneously active links increases the number of infected nodes at each time step, leading to a higher percentage of nodes at a low infection distance from the source since all nodes will be infected before higher infection distances dominate the infection process. Decreasing w will not increase the average temporal path length ad infinitum as a limit is reached when maximally one link is active at each time step. Any further lowering of w will not change the temporal path length, only the duration of the temporal path between nodes will continue to increase. This limit distribution coincides with the distribution of infection-path lengths in continuous time. In continuous time processes, at any precise time instant only one node gets infected as well. Therefore, in this limit the continuous time approximation is valid.

7.2.2 Influence of link density



(a) (b)

Figure 7.5: (a) Distribution of the infection-path length in continuous time simulated on static random graphs with different link density p and 500 nodes. The solution of the SI model is plotted as a reference. (b) Dependence of the average infection-path length on the link density p. The static distance $\ln N/\ln(Np)$ is plotted as a reference.

The homogeneous mixing case is quite artificial. It can only be applied for densely connected communities. Usually, not everyone is connected with everyone else, and contact networks are rather sparse. The average distance between two nodes in graphs with small world properties is of the order of $\ln N$. More specifically, for random graphs it is proportional to $\frac{\ln N}{\ln k}$ [28]. If the random network becomes sparse, the average inter-vertex distance therefore depends on the link density.

We will study the effect of link density in random graphs on the dynamic path lengths. In the case of discrete dynamics for $w = 1$, the temporal path length follows the distance on static graphs. We will focus on the case of continuous dynamics, which corresponds to sparse link activity.

By varying the link density p of the random graph, we regulate the average degree $\langle k \rangle = Np$ and the average inter-vertex distance on the static network $\frac{\ln N}{\ln(Np)}$. Fig. 7.5 shows the distribution of infection-path lengths for random graphs with various link densities. For a large range of link densities, from $p = 1$ to about $p = 0.1$ the average infection-path length stays robust. Here the approximation through infection path length on the homogeneous mixing SI-model is still comparatively good. The slight increase of the average inter-vertex distance on the static graph in this range of link densities does not have any effect on the infection-path length. Random graphs start to become disconnected for $p = \frac{\ln N}{N}$ [30]. When this was the case, we repeated the creation of the random graph up to 200 times, until a connected graph was obtained. Thus, the selection of the random graph used introduced a small bias on graph properties like the degree distribution. Only when the graph becomes sparse and approaches the bond percolation threshold $p = 1/N$ [30], when the average inter-vertex distance increases strongly, it has an influence on the infection-path length.

The distances on the static graph give a constraint on the infection-path length. The number of nodes with an infection-path distance greater than d from the source has a lower limit given by the number of nodes at a distance greater than d on the static graph. The probability for a node on the static random graph to be at a distance greater than d from the source can be roughly approximated by $F_d = \exp(-\frac{1}{N}(Np)^d)$ [13]. The number of nodes which can be reached

p \ w	0.1	0.01	0.001	0.0001
0.1	1	1	0.922	0.226
0.01	1	0.990	0.372	0.045
0.001	0.999	0.501	0.065	-

Table 7.1: For the generated networks with negative binomial weight distribution, average weight w and parameter $p = m/v = Tw/v$, the link density is given, corresponding to the percentage of non-zero weights.

within d transmission steps has an upper limit given by the number of nodes which are at a distance of less than or equal to d steps from the source on the static graph ($N(1 - F_d)$).

By decreasing p, this limit is first reached for the direct neighbors of the source node. For the number of nodes with an infection path distance of $d = 1$ from the source, the upper limit is Np. Looking at the approximation of the distribution of the infection-path lengths using the SI model on a fully connected graph, this limit is attained for $p \geq \ln(N - 1)/N$. However, the approximation using the fully connected graph is already poor for much higher p. A stronger effect will most likely come from the fact that the number of susceptible nodes in Eq. 7.1 is also limited by $1 - F_d$, limiting the increase of infected nodes at low transmission distances. Furthermore, when the graph becomes spares but stays connected, the number of cycles decreases and thus the infection-path length will become more similar to the shortest path length.

7.2.3 Influence of the weight distribution

In contact networks not all connections happen at the same frequency. Some links are stronger, and neighbors will frequent each other more often; some links are weaker, and neighbors will interact rarely. In order to see if different link activities will have an effect on the temporal path length, we create random networks with a negative binomial weight distribution. The negative binomial distribution was chosen because in spite of being sufficiently broad it still has an easily controllable expectation value and variance. The negative binomial distribution is generated with the probability parameter p and the dispersion parameter $n = m^2/(v - m)$, where m is the mean contact time and v the variance. It models the distribution of total contact times and was divided by $T = 10000$ in order to obtain the distribution of weights. We chose the mean contact time m from 1 time unit up to 1000 out of 10000, which corresponds to a weight between $w = 10^{-4}$ and $w = 0.1$, and we chose the variance of the weight distribution v so that the parameter $p = m/v$ varies between 0.1 and 0.001. The weights, including zero-weights, are randomly distributed on the links. The average weight is calculated over all links, including those with weight zero. In Tab. 7.1 the link density is shown for the chosen parameters. In the case of $w = 0.0001$ and $p = 0.001$, the network consisted of more than one connected component. For this case no simulations were done. The static networks were then transformed into dynamic networks where each link is active at any given time with a probability corresponding to its weight w on the static network. Links do not have weights on the dynamic network. The temporal network was generated for a length of $T = 10000$ timesteps.

In Fig. 7.6 the distributions of the temporal path lengths are plotted for different weight distributions. Results are averaged over 100 runs. The main effect on the distribution of temporal path lengths is caused by the average probability of links to be active, which is given by the average weight, similar to (Fig. 7.4). For densely connected networks, increasing the heterogeneity of the weight distribution does not show any effect on the temporal path-length distribution. Only for link densities below 10% does a higher heterogeneity of the weights result in a divergence of

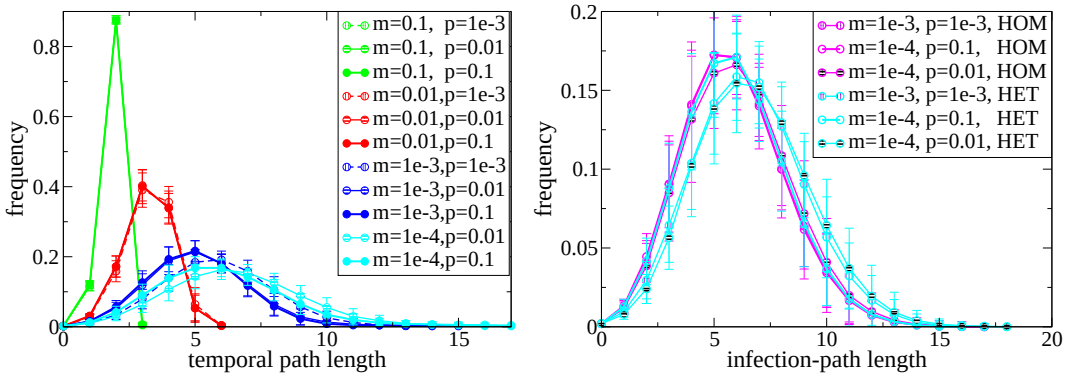


Figure 7.6: Left: distribution of temporal path length in discrete time for various weights w, which are distributed according to a negative binomial distribution with mean weight w and variance $v_w = w/(Tp)$. The weights mark the probability a given link is active at a given time. The weight distribution has a higher variance for smaller p. Right: distribution of infection-path length in continuous time for three of the networks with lowest average weight. For each of these networks, the simulation on an identical network with homogeneous weight (HOM) is added as comparison.

the distribution of temporal path lengths towards longer temporal paths.

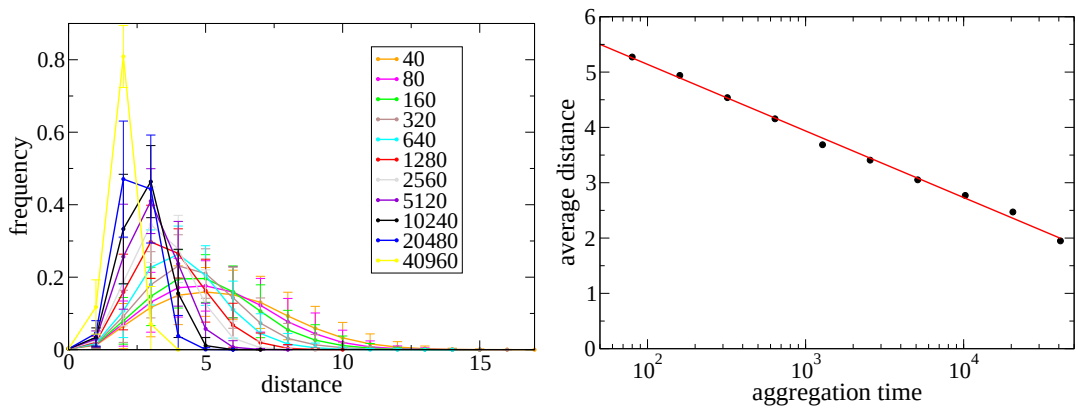
To exclude the possibility that the change in temporal path length is only due to the reduced link density, the simulation was redone for two topologically identical networks with the same average weight, which only differ in the weight distribution. One has a negative binomial weight distribution, and the other one has the same topology but all non-zero weights are replaced by their average. We call them HET and HOM. The simulation was done in the continuous time case in order to better compare between networks with different average weight. The effect that infection-path lengths increase for weight distributions with higher variance in networks with low link density persists.

As long as the network has high link density, there is no influence of the weight distribution. When the network approaches the percolation threshold, then the average transmission path length increases rapidly (see previous section). Links with very low weight are only very rarely active so that the network has an even lower link density most of the time. As the effect of link density is quite strong close to the percolation threshold, even a partial reduction of link density in time can influence the transmission path length. Furthermore it would be interesting to see the effect of topology, like clustering, scale-freeness or different degree distributions, on the transmission path length.

7.3 Distance on face-to-face contact networks

For transmission processes on contact networks, the topology of the network and the burstiness of the dynamics play important roles [44]. In some networks, the dynamics and structure can lead to an optimized transmission, while others are constructed in a way to hinder the flow of information [82, 63]. Concerning the time needed to get from one node to another, Pan et al. [73] have investigated the duration of temporal paths on different networks and found significant differences between the time needed to reach any node on air transport networks or communication networks following temporal paths.

Here we look at the length of temporal paths on face-to-face contact networks. In order to differentiate between temporal effects, like time resolution of the data, time ordering and burstiness in the dynamic network and effects due to the difference in contact frequency between neighboring nodes, we will compare results with two model networks.



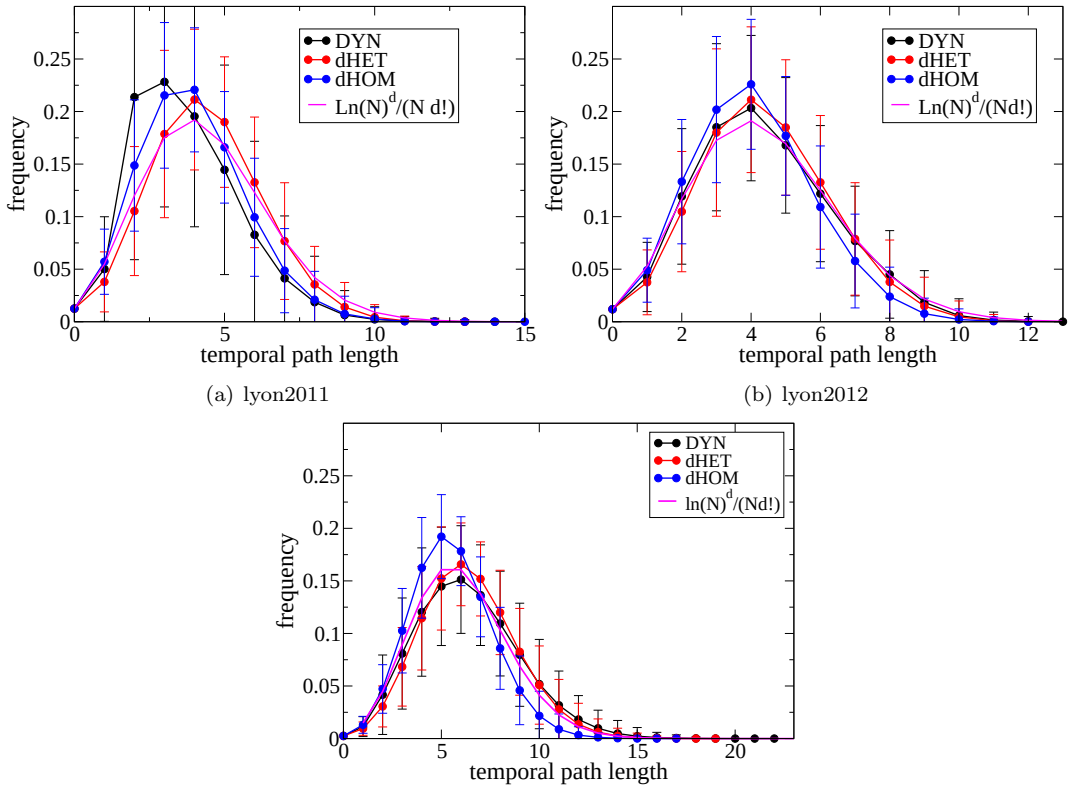
(a) (b)

Figure 7.7: (a) Distribution of temporal path length for the "sfhh" contact network at different aggregation levels. The different colors correspond to different aggregation time steps (given in seconds). (b) Average temporal path length vs. aggregation level.

By aggregating the dynamic contact network over different time lengths, we can control for the time resolution of the underlying data. In Fig. 7.7 the distribution of the temporal path length on the "sfhh" dataset is shown for different time resolutions. Similar to Fig. 7.4 where links were active with probability w so that more links were active at the same time for higher w, here, more links are active at the same time due to the higher degree of the network snapshots with higher aggregation time. This higher average instantaneous degree of the network leads to a shorter average temporal path length. The average temporal path length depends logarithmically on the time resolution in the given range of aggregation times. However, the link density in each network snapshot cannot increase infinitely with aggregation time. Thus, also the temporal path length will eventually stop decreasing. Also, even if the time resolution was ever more precise, it is not likely that the temporal path length continues to increase.

The effect of the different dynamic properties on the temporal distances can be seen in Fig. 7.8 for the datasets "sfhh", "lyon2011" and "lyon2012". In order to control for the effect of the network dynamics, we create two networks with random dynamics, dHET and dHOM, for each data set using the static networks HET and HOM. The links in dHET and dHOM are active with a probability according to their weight on the static networks HET and HOM. Thus, dHET is a temporal network with Poissonian contact-time distribution and no burstiness, but the same aggregated weights as the corresponding temporal network (DYN). The dHOM network also has Poissonian event dynamics, but all non-zero links have the same probability to be active. Results for the dHET and dHOM network are averaged over different realizations of the networks. Results for the DYN network are averaged over different starting times. The bursty dynamics leads to slightly shorter average temporal path length, compared to the networks with random dynamics. The weight distribution in dHET leads to slightly longer temporal paths compared to dHOM, in agreement with Sec. 7.2.3.

At first it seems surprising that the temporal path length for the DYN network for "lyon2011"



(a) lyon2011

(b) lyon2012

(c) sfhh

Figure 7.8: Temporal path-length distribution for the contact networks of the "sfhh", "lyon2011" and "lyon2012" data sets. The temporal path length of two model networks, dHET and dHOM, is also shown. In dHET contacts have the same probability as in the original data, but are random in time. In dHOM all contacts have the same probability so that the average activity of dHOM is identical to dHET and the original data.

is lower than for the networks with Poissonian dynamics. Just as in the random dynamics case dHET, some links are much less often active than others. However, while the temporarily reduced link density increases the temporal path length in the random dynamic model, dHET, the opposite effect is witnessed on the original data set. One reason for shorter temporal path length might be the fact that temporal correlations and burstiness lead to temporarily smaller networks with higher link density [6]. The distribution of temporal path lengths for the DYN network will most likely also depend on the activity fluctuations and the starting time. This remains to be tested. The continuous approximation of the infection-path length distribution, $\ln(N)^d/d!$, can also give a first estimate for the temporal path length distribution. It captures the essence of the distance on the dynamic graph better than the static distance does, but is obtained for Poisson dynamics with low link activity on fully connected networks. Temporal properties like correlations between links or activity fluctuations, which can lead to lower temporal path lengths, or sparse connectivity, which can lead to higher temporal path lengths, will worsen the approximation. Especially when the underlying network becomes very sparse, or when the dynamics shows a high number of links which are active at the same time, the approximation loses its validity.

7.4 Conclusion

As could be seen for the duration of temporal paths [73] and the length of infection paths [71], also the length of temporal paths between two nodes is much higher than the distance on the static network. We have looked in particular at the distribution of the temporal path lengths and the infection path lengths. Both distributions have similar properties and can be approximated by $p(d) = \ln(N)^d/(Nd!)$. We furthermore tested how different properties of static networks like the link density and heterogeneity of weights influence the infection path lengths and the temporal path lengths. The latter is also influenced by temporal properties like a high instantaneous average degree. It remains to be tested how other topological and temporal properties influence these distributions. The distribution of infection path lengths for SIR models on temporal networks would also be of interest as, depending on the parameters of the SIR model, the paths the infection takes are often longer than the temporal paths. Preliminary results which are not shown here suggest that the infection path lengths on temporal networks increases on average when the epidemic is slower up to the point where not all nodes are reached anymore.

.

Chapter 8

Conclusions

With the advent of new technologies, contact data has become available with higher details and to larger extent. This abundance of detailed data opens up new opportunities to model social interactions and spreading of information or epidemics in social settings with unprecedented precision.

However, with the high availability of data at various levels of detail rises also the question of how to best represent data for modeling. Data is always a partial representation of reality, taken in order to answer specific questions. It is not evident how much data is needed and which amount of detail is optimal to best answer these questions. The problem we have put our focus on is the spread of epidemics. In order to tackle this problem, we have used face-to-face contact data, as face-to-face contacts can be a valuable proxy for the spread of influenza-like diseases. A closer look at the data sets has revealed temporal patterns on different time scales, like weekly patterns, day-night patterns or daily patterns. Furthermore, at any time new nodes are introduced in the data and the time between contacts can have any timescale.

In a previous work, Stehlé et al. [88] have shown that under certain conditions static networks can be adequate representations of temporal networks in epidemic modeling. We investigated the role of the model parameter sets and the time over which the static network was aggregated in order to assess the conditions under which static networks can sufficiently represent temporal networks. We have found that for our temporal data sets, which have high temporal variability and an introduction of new nodes at any time scales, spreading on a static network representation can be a good approximation for spread on a temporal network if the length of the epidemic matches the time over which the network was aggregated. Otherwise, the additional links and nodes which the static network accumulates with longer aggregation times can lead to an overestimation of the outcome of the epidemic. Similarly, if the epidemic duration is longer than the aggregation time of the static network, the outcome of the epidemic can be severely underestimated. This is also true for spreading on a dynamic network when the data is repeated. Here as well, the introduction of new nodes and links after the end of the data set would have altered the outcome of the epidemic. In the light of the temporal limits of data sets not only epidemic simulations on static networks need to be critically reassessed depending on the parameters used but also epidemic simulations on repeated temporal data sets.

In order to obtain a more accurate result for the distribution of final sizes of the epidemic, instead of fully aggregated networks, partly aggregated networks, for example daily networks (see Stehlé et al. [88]), are a viable alternative. Investigating on the necessary level of detail in the temporal resolution of the data, we looked at the influence of temporal patterns of the network on the spreading process. The influence of the variability of the data on the outcome of

the epidemic was stronger for faster epidemics, which had high probability of propagation and recovery. While for slowly developing epidemics with low β and μ daily fluctuations played only a minor role, fast processes were strongly influenced by the temporal structure of the network at short timescales. We found that the question whether or not the data can be simplified depends as much on the patterns of the data as on the process on the data. Indeed, the temporal resolution could be reduced to the order of the infectious period of nodes. We could thus shed some light on the question at which resolution data is accurate enough for a specific epidemic process.

This estimate for the maximum aggregation time however only concerns the exact ordering and timing of contacts, the average time a link is active over the aggregation period still needs to be measured. Increasing the aggregation time steps by using less accurate measuring procedures without keeping information on the aggregated activity of contact links will otherwise overestimate the total time individuals spent in contact over a certain duration, altering the outcome of the epidemic in a non-negligible way. The optimal choice of a minimal time step up to which contacts can be aggregated without keeping information on average contact activity seems nontrivial and depends on the measuring procedure.

In the quest between accuracy of the outcome of epidemic simulations and practicability as well as generalizability of the data representations, we went one step further, introducing a data representation which lies in between the contact matrices and the static network. The contact matrix representation is much used in epidemiology as it is a big improvement over the homogeneous mixing hypothesis but still only needs a minimum amount of information. Contact matrices only require the average time different groups spend in contact with each other. This representation is highly unspecific and independent of the number of individuals in each group. However, it completely ignores the heterogeneity of total contact time spent between members of different groups or among members of the same group. The data representation we introduced is a contact matrix of distributions and thus easier to generalize and to transport into other settings than the individual based exact static network, but at the same time it keeps information on the heterogeneity of the link weights. To optimally use the advantage this method brings, the choice of the right groups is important. As the information on the distribution of total contact time is largely kept, it is of particular importance to choose groups in such a way that the structural properties of the network are maintained. Grouping nodes according to their degree, so that nodes with similar degree on the static network are placed in the same group, is a promising first approach. The natural groups, given by the roles the individuals played in the hospital, proved to be a better choice though. It is not yet very clear which properties of the nodes are good proxies to reach an optimal choice of groups. Possibly, community algorithms can work to find groups which keep much of the heterogeneity of the network structure. This remains to be tested. Furthermore, it would be interesting to find the optimal number of groups into which the network can be partitioned, so that on the one hand the information that is needed to construct the network is minimized and on the other hand the outcome of the epidemic simulation remains reasonably accurate. However, independent of the choice of groups, the contact matrix of distributions gives much more reliable results on the outcome of the epidemic than the contact matrix representation. Being just as generalizable, we hope that it can eventually replace contact matrix representations in large multi-scale epidemic models, thus increasing the accuracy of the outcome of simulations.

We tested the contact matrix of distributions with respect to its ability to function as a versatile representation that can be applied to different situations and with respect to its ability to suggest immunization strategies. In an attempt to use the contact matrix of distributions for different but similar settings, we tested it on three different data sets, two of which were obtained with the same participants at the same venue. The contact matrix allowed us to adapt

the number of individuals in each group in order to provide a prediction for the epidemic spread on a different data set with different numbers of individuals per group. We thus could obtain an idea of the accuracy of predictions which can be made on a future epidemic outbreak when using available data from the same or similar settings. The data sets which were one year apart could show the limits of predictions in general if data is used as a prediction for a different situation. At the same time, when the contact matrix of distributions was obtained on the first part of a long data set, predictions for the second part based on the first part were fairly accurate. It would be interesting to test several longer data sets in order to better estimate the variability of the corresponding contact matrices in order to assign a confidence interval to the respective simulations depending on the predicted variations of the data.

This can also be important for the design of immunization strategies, as those are also based on data from the past but applied to an unknown situation in the future. However, in order to understand which nodes need to be vaccinated, it is useful to create and apply immunization strategies on the same data set. We found that by keeping information on the distribution of the total amount of time spent between two people, the contact matrix of distributions can also better inform immunization strategies than the contact matrix, as the former preserves the information of the average degree per group, while the latter only contains information on the average time spent between individuals. Immunization strategies relying on an ordering based on the average degree of groups proved to be nearly as good as immunization strategies based on the individual degree of nodes. Furthermore, when reducing the size of datasets, group based immunization strategies were more robust than strategies based on individual nodes. This is mainly due to the high variability of the importance of nodes as spreader of information or diseases over time. We could confirm the existence of this variability in our data sets. Therefore the usefulness of individual based immunization strategies is limited. At the same time, we saw that even a small amount of data will already lead to reasonably good degree-based immunization strategies, which will only improve slowly with the information given by longer datasets.

The changing importance of nodes over time poses problems when looking for optimal immunization strategies that use temporal information of the data. We have conceived a measure which is able to estimate the significance of a node by considering its direct influence on the temporal paths of an SI process. This significance fluctuates with the starting time of the SI process. In our data sets, different nodes were significant for the epidemic process at different times. Thus, when averaging over those times in order to find a global list for immunization, much of the additional temporal information was lost, so that the resulting immunization strategy is not better than a strategy conceived on a static network. Possibly, this method can lead to better results when temporal patterns repeatedly lead to the significance of some nodes but not of others. However, the method to calculate the significance used here was based on temporal paths on the network. Often epidemics do not follow the temporal path. Especially if the propagation probability is low, the epidemic will follow longer paths on the network. It could be useful to conceive a measure which takes the multitude of possible transmission paths into account and also their dependence on the spreading parameters. Also, in the same way as Takaguchi et al. [95] have tested the importance of events on the temporal network, it might be a good idea not to look for a global immunization strategy which is valid for the entire temporal network but instead test the efficiency of the removal of nodes at certain time periods, similar to applying protective measures against epidemic spread, like masks doctors wear in hospitals. It could be interesting to furthermore test other time-varying properties, like the partially aggregated degree, in order to predict the importance of nodes at precise times, similar to the degree on a static network.

As a side note, we also looked at the distribution of temporal path length for dynamic networks. We compared it to the distribution of infection path lengths on the corresponding aggregated network and could show that it is reasonably well approximated by the latter, which

again can be approximated by a simple solution of the SI model's differential equations.

Overall in this thesis we have contributed to provide insights into questions concerning dynamic epidemic processes on data-driven, temporal networks. In particular, we have investigated the influence of data representations on the outcome of epidemic processes, shedding some light on the question how much detail is necessary for the data representation and its dependence on the spreading parameters. By introducing an improvement to the contact matrix representation we hope we could provide a data representation that could in the future be integrated into multi-scale epidemic models in order to improve the accuracy of predictions and corresponding immunization strategies. We could also point out some of the ways dynamic processes are influenced by temporal properties of the data. However, much remains to be done. Especially, a quantification of the results on the basis of models might be useful.

Appendix A

Appendix

	Data				
	A	D	N	P	C
A	298	1.16	24.7	0.95	1.92
D	1.16	20.8	3.99	0.95	1.20
N	24.7	3.99	47.3	2.32	2.57
P	0.95	0.95	2.32	1.27	46.9
C	1.92	1.20	2.57	46.9	1.80

Table A.1: Average contact time in seconds per day between two members of each group for the "obg" data set.

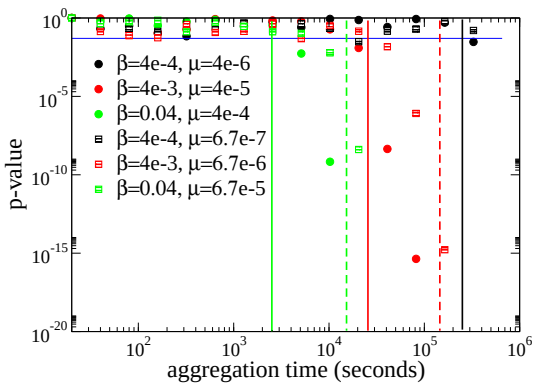
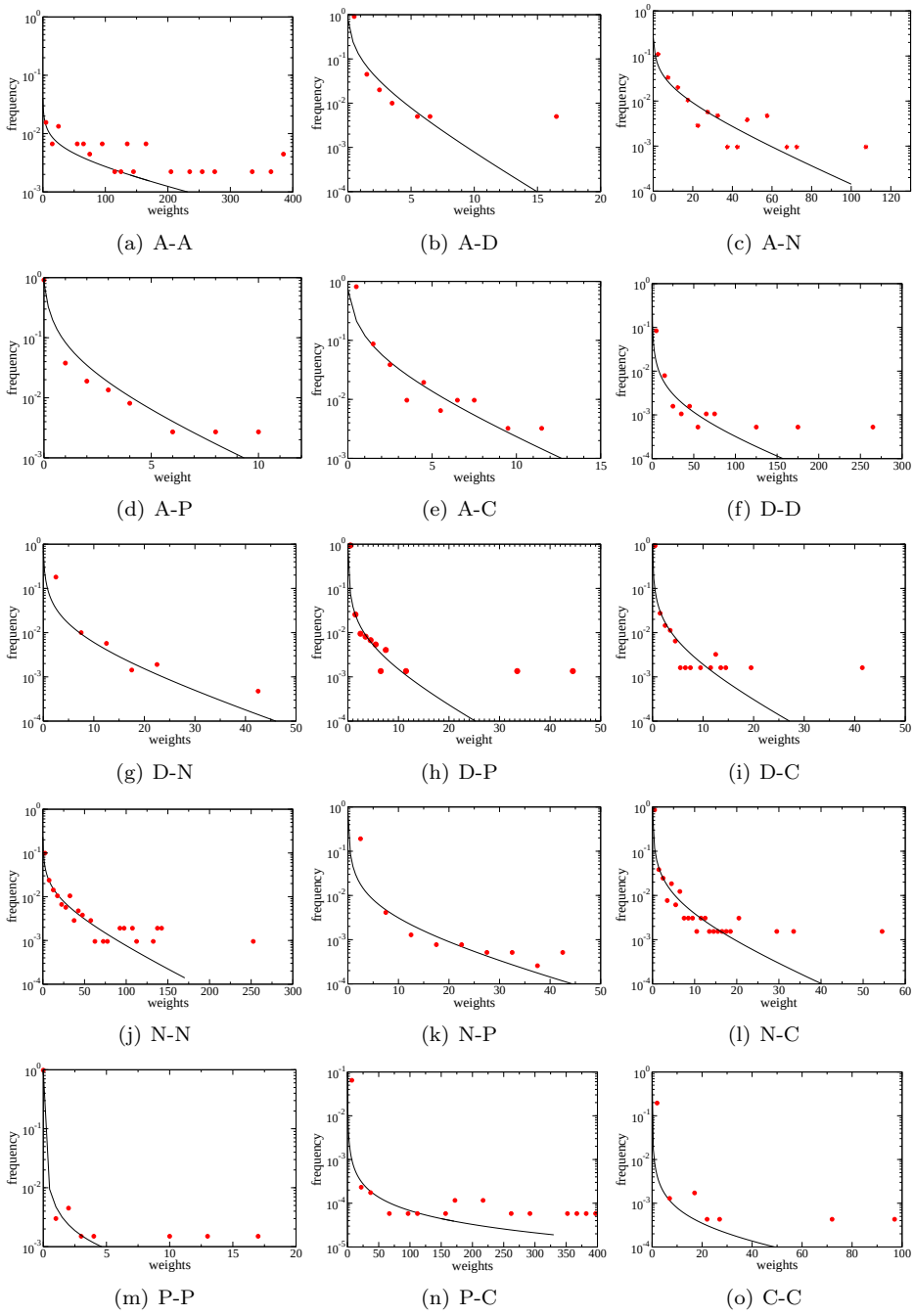


Figure A.1: The χ^2-test has been used to compare the distribution of final cases for a partially aggregated network with given aggregation timestep and the original temporal network. The p-value is calculated as described in Numerical Recipes [80] (p.733), the number of bins was chosen as the smallest set of bins which were non-zero for both distributions. The vertical lines correspond to the value of μ^{-1}. Continuous lines correspond to the value used for simulations marked by a circle with the same color, dashed lines correspond to the value used by simulations marked by a square of the same color.



(a) A-A (b) A-D (c) A-N

(d) A-P (e) A-C (f) D-D

(g) D-N (h) D-P (i) D-C

(j) N-N (k) N-P (l) N-C

(m) P-P (n) P-C (o) C-C

Figure A.2: Weight distribution for the weights between and among different groups in the "obg" data set, where A-Assistants, D-Doctors, N-Nurses, P-Patients, C-Caregivers. The weight is given in seconds per day. The maximum likelihood fit with a negative binomial distribution is shown as well. The parameters of the fit are given in Tab. A.2 and Tab. A.3.

	Fit				
	A	D	N	P	C
A	300(60)	1.2(0.2)	25(3)	0.95(0.16)	2.0(0.3)
D	1.2(0.2)	21(5)	4.0(0.6)	0.95(0.17)	1.2(0.2)
N	25(3)	4.0(0.6)	47(5)	2.3(0.4)	2.6(0.4)
P	0.95(0.16)	0.95(0.17)	2.3(0.4)	1.3(0.7)	47(16)
C	2.0(0.3)	1.2(0.2)	2.6(0.4)	47(16)	1.8(0.9)

Table A.2: Average contact time in seconds per day between two members of each group from the fit with a negative binomial distribution on the "obg"data set. The numbers in parenthesis are the standard errors as given by R's "fitdistr" function.

	Fit				
	A	D	N	P	C
A	0.615(0.014)	0.195(0.002)	0.404(0.0018)	0.136(0.0007)	0.215(0.0013)
D	0.195(0.002)	0.112(2.10-4)	0.1278(2.10-4)	0.0482(5.10-5)	0.0602(8.10-5)
N	0.404(0.0018)	0.1278(2.10-4)	0.3696(0.0013)	0.05652(4.10-5)	0.0845(9.10-5)
P	0.136(7.10-4)	0.0482(5.10-5)	0.0565(4.10-5)	0.00489(1.8.10-6)	0.00718(9.10-7)
C	0.215(0.0013)	0.0602(8.10-5)	0.0845(9.10-5)	0.00718(9.10-7)	0.009(6. 10-6)

Table A.3: The r-parameter obtained from fits with a negative binomial distribution on the "obg"data set. The numbers in parenthesis are the standard errors as given by R's "fitdistr" function. The variance of the distribution is given by $m + m^2/r$, where m is the mean of the distribution.

		DYN		HET		HOM		CM		CM0		CMD	
	runs	EP	AR< 10%	EP	AR< 10%	EP	AR< 10%	EO	AR< 10%	EP	AR< 10%	EP	AR< 10%
All	16000	0.60	0.86	0.47	0.80	0.41	0.50	0.50	0.80	0.33	0.48	0.38	0.65
Assistants	1344	0.35	0.51	0.21	0.34	0.20	0.27	0.18	0.26	0.18	0.28	0.20	0.33
Doctors	2690	0.71	0.93	0.65	0.87	0.40	0.49	0.56	0.81	0.57	0.84	0.59	0.91
Nurses	2823	0.50	0.70	0.39	0.56	0.19	0.28	0.33	0.48	0.33	0.52	0.36	0.58
Patients	4975	0.66	0.95	0.51	0.91	0.54	0.64	0.30	0.44	0.40	0.71	0.58	0.92
Caregivers	4168	0.59	0.95	0.45	0.91	0.47	0.58	0.27	0.40	0.35	0.68	0.52	0.90

Table A.4: Fraction of runs that lie under a given threshold for the different network models. For each network model, the dependence on the starting group is also taken into account. Simulations are done for parameter set 1 (see Tab. 4.2).

		DYN		HET		HOM		CM		CM0		CMD	
	runs	EP	AR< 10%	EP	AR< 10%	EP	AR< 10%	EO	AR< 10%	EP	AR< 10%	EP	AR< 10%
All	16000	0.62	0.81	0.35	0.63	0.28	0.32	0.21	0.27	0.26	0.40	0.37	0.59
Assistants	1344	0.44	0.51	0.13	0.16	0.10	0.12	0.10	0.12	0.10	0.12	0.11	0.14
Doctors	2690	0.70	0.84	0.51	0.67	0.27	0.30	0.40	0.55	0.39	0.57	0.42	0.66
Nurses	2823	0.51	0.64	0.26	0.33	0.11	0.13	0.20	0.26	0.21	0.27	0.21	0.30
Patients	4975	0.68	0.90	0.40	0.79	0.39	0.44	0.18	0.23	0.28	0.44	0.47	0.74
Caregivers	4168	0.62	0.90	0.31	0.77	0.32	0.37	0.16	0.20	0.24	0.41	0.42	0.72

Table A.5: Fraction of runs that lie under a given threshold for the different network models. For each network model, the dependence on the starting group is also taken into account. Simulations are done for parameter set 2 (see Tab. 4.2).

Bibliography

[1] Harith Alani, Martin Szomszor, Ciro Cattuto, Wouter Broeck, Gianluca Correndo, and Alain Barrat. Live social semantics. In Abraham Bernstein, DavidR. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009*, volume 5823 of *Lecture Notes in Computer Science*, pages 698–714. Springer Berlin Heidelberg, 2009.

[2] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

[3] Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[4] Paolo Bajardi, Alain Barrat, Fabrizio Natale, Lara Savini, and Vittoria Colizza. Dynamical patterns of cattle trade movements. *PloS one*, 6(5):e19869, 2011.

[5] Duygu Balcan, Bruno Gonalves, Hao Hu, Jos J. Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, 1(3):132 – 145, 2010.

[6] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[7] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[8] Andrea Baronchelli, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06014+, June 2006.

[9] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[10] A. Barrat, M. Barthlemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[11] Alain Barrat, Ciro Cattuto, Vittoria Colizza, Jean-Franois Pinton, Wouter Van den Broeck, and Alessandro Vespignani. High resolution dynamical mapping of social interactions with active rfid. *ArXiv e-prints*, November 2008.

[12] Claude Berge. *Graphs and hypergraphs*, volume 6. North-Holland publishing company Amsterdam, 1973.

[13] Vincent D. Blondel, Jean-Loup Guillaume, Julien M. Hendrickx, and Raphaël M. Jungers. Distance distribution in random graphs and application to network exploration. *Phys. Rev. E*, 76:066101, Dec 2007.

[14] Sally Blower and Myong-Hyun Go. The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? *BMC Medicine*, 9:1–3, 2011.

[15] Dan Braha and Yaneer Bar-yam. From centrality to temporary fame: dynamic centrality in complex networks. *Comlexity*, 12, 2006.

[16] Romulus Breban, Raffaele Vardavas, and Sally Blower. Theory versus data: How to calculate r¡sub¿0¡/sub¿? *PLoS ONE*, 2(3):e282, 03 2007.

[17] Wouter Broeck, Corrado Gioannini, Bruno Goncalves, Marco Quaggiotto, Vittoria Colizza, and Alessandro Vespignani. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infectious Diseases*, 11(1):37, 2011.

[18] Carter T. Butts. Revisiting the foundations of network analysis. *Science*, 325(5939):414–416, 2009.

[19] C. Castellano, D. Vilone, and A. Vespignani. Incomplete ordering of the voter model on small-world networks. *EPL (Europhysics Letters)*, 63(1):153–158, 2003.

[20] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5, 2010.

[21] Marta Luisa Ciofi degli Atti, Stefano Merler, Caterina Rizzo, Marco Ajelli, Marco Massari, Piero Manfredi, Cesare Furlanello, Gianpaolo Scalia Tomba, and Mimmo Iannelli. Mitigation measures for pandemic influenza in italy: An individual based model considering different scenarios. *PLoS ONE*, 3(3):e1790, 03 2008.

[22] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[23] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4628, Nov 2000.

[24] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Phys. Rev. Lett.*, 86:3682–3685, Apr 2001.

[25] Reuven Cohen, Shlomo Havlin, and Daniel ben Avraham. Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.*, 91:247901, Dec 2003.

[26] Elizabeth Costenbader and Thomas W Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283 – 307, 2003.

[27] Luca Dall'Asta, Andrea Baronchelli, Alain Barrat, and Vittorio Loreto. Nonequilibrium dynamics of language games on complex networks. *Phys. Rev. E*, 74:036105, Sep 2006.

[28] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Metric structure of random networks. *Nuclear Physics B*, 653(3):307 – 338, 2003.

[29] Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, and Edmunds WJ. Using time-use data to parameterize models for the spread of close-contact infectious diseases. *Am J Epidemiol*, 168:1082–90, 2008.

[30] P. Erdös and A Rényi. On the evolution of random graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960.

[31] Scott L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):pp. 1464–1477, 1991.

[32] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[33] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[34] C. Godrèche, H. Grandclaude, and J.M. Luck. Finite-time fluctuations in the degree statistics ofgrowing networks. *Journal of Statistical Physics*, 137(5-6):1117–1146, 2009.

[35] E. Goldstein, A. Apolloni, B. Lewis, J. C. Miller, M. Macauley, S. Eubank, M. Lipsitch, and J. Wallinga. Distribution of vaccine/antivirals and the least spread line in a stratified population. *Journal of The Royal Society Interface*, 7(46):755–764, 2010.

[36] E. Goldstein, K. Paur, C. Fraser, E. Kenah, J. Wallinga, and M. Lipsitch. Reproductive numbers, epidemic spread and control in a community of households. *Mathematical Biosciences*, 221(1):11 – 25, 2009.

[37] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[38] Caroline Breese Hall. The spread of influenza and other respiratory viruses: Complexities and conjectures. *Clinical Infectious Diseases*, 45(3):353–359, 2007.

[39] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103:038702, Jul 2009.

[40] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, and A.E. Tozzi. Close encounters in a pediatric ward: measuring face-toface proximity and mixing patterns with wearable sensors. *PLoS ONE*, 6(2): e17144, 2011.

[41] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-Franois Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166 – 180, 2011.

[42] Mossong J, AND Jit M Hens N, Beutels P, Auranen K, Mikolajczyk R, Massari M, Salmaso S, Tomba GS, Wallinga J, Heijne J, Sadkowska-Todys M, Rosinska M, and Edmunds WJ. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5:e74, 2008.

[43] Wallinga J, Teunis P, and Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol*, 164:936–44, 2006.

[44] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E*, 83:025102, Feb 2011.

[45] Matt Keeling. The implications of network structure for epidemic dynamics. *Theoretical Population Biology*, 67(1):1 – 8, 2005.

[46] Matt J Keeling and Ken T.D Eames. Networks and epidemic models. *Journal of The Royal Society Interface*, 2(4):295–307, 2005.

[47] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.

[48] Gueorgi Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3):247 – 268, 2006.

[49] Eames KTD. Modelling disease spread through random and regular contacts in clustered populations. *Theor Popul Biol*, 73:104–11, 2008.

[50] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73:016102, Jan 2006.

[51] Sungmin Lee, Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Exploiting temporal network structures of human interaction to effectively immunize populations. *PLoS ONE*, 7(5):e36439, 05 2012.

[52] Fredrik Liljeros, Christofer R. Edling, Luis A. Nunes Amaral, H. Eugene Stanley, and Yvonne Aberg. The web of human sexual contacts. *Nature*, 411:907– 908, 2001.

[53] William G. Lindsley, William P. King, Robert E. Thewlis, Jeffrey S. Reynolds, Kedar Panday, Gang Cao, and Jonathan V. Szalajda. Dispersion and exposure to a cough-generated aerosol in a simulated medical examination room. *Journal of Occupational and Environmental Hygiene*, 9(12):681–690, 2012. PMID: 23033849.

[54] Alun L. Lloyd and Robert M. May. How viruses spread among computers and people. *Science*, 292(5520):1316–1317, 2001.

[55] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438:355–9, 2005.

[56] Halloran ME, Ferguson NM, Eubank S, Longini IM Jr, Cummings DA, Lewis B, Xu S, Fraser C, Vullikanti A, Germann TC, Wagener D, Beckman R, Kadau K, Barrett C, Macken CA, Burke DS, and Cooley P. Modeling targeted layered containment of an influenza pandemic in the united states. *Proc Natl Acad Sci U S A*, 105:4639–44, 2008.

[57] Jan Medlock and Alison P. Galvani. Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708, 2009.

[58] R.T. Mikolajczyk, M.K. Akmatov, S. Rastin, and M. Kretzschmar. Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology & Infection*, 136:813–822, 5 2008.

[59] Stanley Milgram. The small-world problem. *Psychology Today*, 1(1):61–67, 1967.

[60] Joel C Miller. Spread of infectious disease through clustered populations. *Journal of the Royal Society Interface*, 6(41):1121–1134, 2009.

[61] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[62] Byungjoon Min, K.-I. Goh, and I.-M. Kim. Absence of epidemic outbreaks with heavy-tailed contact dynamics. 2013.

[63] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Phys. Rev. E*, 83:045102, Apr 2011.

[64] Keeling MJ. The effects of local spatial structure on epidemiological invasions. *Proc Biol Sci*, 266:859–67, 1999.

[65] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 26(4):521–529, 2002.

[66] Hens N, Goeyvaerts N, Aerts M, Shkedy Z, Van Damme P, and Beutels P. Mining social mixing patterns for infectious disease models based on a two-day population survey in belgium. *BMC Infect Dis*, 20;9:5, 2009.

[67] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001.

[68] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, Jul 2002.

[69] Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, 1992.

[70] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[71] Jukka P. Onnela and Nicholas A. Christakis. Spreading paths in partially observed social networks. *Physical Review E*, 85:036106+, March 2012.

[72] Beutels P, Shkedy Z, Aerts M, and Van Damme P. Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiol Infect*, 134:1158–66, 2006.

[73] R.K. Pan and J. Saramki. Path lengths, correlations, and centrality in temporal networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 84(1 Pt 2):016105, 2011.

[74] André Panisson, Alain Barrat, Ciro Cattuto, Wouter Van den Broeck, Giancarlo Ruffo, and Rossano Schifanella. On the dynamics of human proximity for data diffusion in ad-hoc networks. *Ad Hoc Networks*, 10(8):1532–1543, 2012.

[75] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, Apr 2001.

[76] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics in finite size scale-free networks. *Physical Review E*, 65(3):035108, 2002.

[77] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Phys. Rev. E*, 65:036104, Feb 2002.

[78] N. Perra, B. Gonalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Sci. Rep.*, 2, 2012.

[79] Philip M. Polgreen, Troy Leo Tassier, Sriram Venkata Pemmaraju, and Alberto Maria Segre. Prioritizing healthcare worker vaccinations on the basis of social network analysis. *Infect Control Hosp Epidemiol*, 31(9):893900, 2010.

[80] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[81] Jonathan M Read, Ken T.D Eames, and W. John Edmunds. Dynamic social networks and the implications for the spread of infectious disease. *Journal of The Royal Society Interface*, 5(26):1001–1007, 2008.

[82] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*, 7(3):e1001109, 03 2011.

[83] Marcel Salathé and James H. Jones. Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol*, 6(4):e1000736, 04 2010.

[84] sociopatterns. hypertext2009 data. http://www.sociopatterns.org/datasets/hypertext-2009-dynamic-contact-network/. 2012.

[85] sociopatterns. Sociopatterns project. http://www.sociopatterns.org. 2012.

[86] Michele Starnini, Anna Machens, Ciro Cattuto, Alain Barrat, and Romualdo Pastor-Satorras. Immunization strategies for epidemic processes in time-varying contact networks. *Journal of Theoretical Biology*, 2013.

[87] Luc Steels. A self-organizing spatial vocabulary. *Artificial Life*, 2:319–332, 1995.

[88] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Rgis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, and P. Vanhems. Simulation of a seir infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9:87, 2011.

[89] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Rgis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.

[90] Juliette Stehlé, Alain Barrat, and Ginestra Bianconi. Dynamical and bursty interactions in social networks. *Phys. Rev. E*, 81:035101, Mar 2010.

[91] Michael P. H. Stumpf and Carsten Wiuf. Sampling properties of random graphs: The degree distribution. *Phys. Rev. E*, 72:036118, Sep 2005.

[92] Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224, 2005.

[93] Del Valle SY, Hyman JM, Hethcote HW, and et al. Mixing patterns between age groups in social networks. *Soc Networks*, 29:539–554, 2007.

[94] Smieszek T, Fiebig L, and Scholz RW. Models of epidemics: when contact repetition and clustering should be included. *Theor Biol Med Model*, 6:11, 2009.

[95] Taro Takaguchi, Nobuo Sato, Kazuo Yano, and Naoki Masuda. Importance of individual events in temporal networks. *New Journal of Physics*, 14(9):093003, 2012.

[96] John Tang, Mirco Musolesi, Cecilia Mascolo, and Vito Latora. Characterising temporal distance and reachability in mobile and online social networks. *SIGCOMM Comput. Commun. Rev.*, 40(1):118–124, January 2010.

[97] Germann TC, Kadau K, Longini IM Jr, and Macken CA. Mitigation strategies for pandemic influenza in the united states. *Proc Natl Acad Sci U S A*, 103:5935–40, 2006.

[98] R. Tellier. Review of aerosol transmission of influenza a virus. *Emerg. Infect. Dis.*, Nov 2006.

[99] Colizza V, Barrat A, Barthelemy M, and Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemic. *Proc Natl Acad Sci*, 103:2015–20, 2006.

[100] W. Van den Broeck, C. Cattuto, A. Barrat, M. Szomszor, G. Correndo, and H. Alani. The live social semantics application: a platform for integrating face-to-face presence with on-line social networking. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pages 226–231, 2010.

[101] Alexei Vazquez, Balázs Rácz, András Lukács, and Albert-László Barabási. Impact of non-poissonian activity patterns on spreading processes. *Phys. Rev. Lett.*, 98:158702, Apr 2007.

[102] Erik Volz and Lauren Ancel Meyers. Susceptible infected recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society B: Biological Sciences*, 274(1628):2925–2934, 2007.

[103] Erik Volz and Lauren Ancel Meyers. Epidemic thresholds in dynamic contact networks. *Journal of The Royal Society Interface*, 6(32):233–241, 2009.

[104] Erik M Volz, Joel C Miller, Alison Galvani, and Lauren Ancel Meyers. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS computational biology*, 7(6):e1002042, 2011.

[105] Jacco Wallinga, Michiel van Boven, and Marc Lipsitch. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences*, 107(2):923–928, 2010.

[106] Duncan J. Watts. Networks, dynamics and the small-world phenomenon. *American Journal of Sociology*, 105(2):493–527, 1999.

[107] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[108] Thomas P. Weber and Nikolaos I. Stilianakis. Inactivation of influenza a viruses in the environment and modes of transmission: A critical review. *Journal of Infection*, 57(5):361 – 373, 2008.

[109] Edmunds WJ, O'Callaghan CJ, and Nokes DJ. Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proc Biol Sci*, 264:949–57, 1997.