CCEGN5 workshop program

Monday, June 6th

Richard Kenyon	Graphons and tilings	10:00-11:00
Coffee		
Yufei Zhao	Equiangular lines and eigenvalue multiplicities	11:30-12:30
Lunch		
Ginestra Bianconi	The dynamics of higher-order networks: the effect of topology and triadic interactions	3:00-4:00
Coffee		
Filippo Radicchi	Identifying the optimal embedding dimension of networks	4:30-5:30

Tuesday, June 7th

Peter Orbanz	Limit theorems for distributions invariant under a group of transformations	10:00-11:00
Coffee		
François Caron	Sparse graphs and multigraphs based on exchangeable random measures	11:30-12:30
Lunch		
Diego Garlaschelli	Multiscale network renormalization: scale-invariance without geometry	3:00-4:00
Coffee		
Giovanni Petri	Between higher-order mechanisms and phenomena	4:30-5:30
Workshop Dinner	Estia, 117 Main St, Falmouth, MA 02540	6:00-8:00

Wednesday, June 8th

Marta Sales-Pardo	Using inference for recommendation and to obtain models from data: Is data always good enough?	10:00-11:00
Coffee		
Clara Stegehuis	Optimal structures in complex networks: where can we find them?	11:30-12:30
Lunch		
Tiago Peixoto	Disentangling homophily, community structure and triadic closure in networks	3:00-4:00
Coffee		
Thomas Bläsius	Theoretical algorithm analysis meets practical data	4:30-5:30

Thursday, June 9th

Marián Boguñá	Topological phase transitions in the geometric configuration model	10:00-11:00
Coffee		
Maksim Kitsak	Finding shortest and nearly shortest path nodes in large substantially incomplete networks	11:30-12:30
Lunch		
Fragkiskos Papadopoulos	Dynamics of cold versus hot random hyperbolic graphs	3:00-4:00
Coffee		
Johannes Lengler	Information spreading in geometric inhomogeneous random graphs	4:30-5:30

Friday, June 10th

Miklós Rácz	Correlated stochastic block models: graph matching and community recovery	10:00-11:00
Coffee		
Subhabrata Sen	Incorporating network side information in statistical learning: a statistical physics based approach	11:30-12:30
Lunch		
Souvik Dhara	Spectral recovery of binary censored block models	3:00-4:00
Coffee		
Paul Krapivsky	Dynamic space packing	4:30-5:30

The talk abstracts are listed in the following pages in the order of the schedule of talks.

Richard Kenyon – *Graphons and tilings*

This is joint work with Cosmin Pohoata (Yale). We define a graphon version of the random tiling model. We compute its entropy and covariance functions. For random tilings of domains in $\mathbb{Z}^d$ we show that the scaling limit satisfies conformal invariance in some settings, for example for tilings of planar domains with translates of the "L" polyomino.

Yufei Zhao – *Equiangular lines and eigenvalue multiplicities*

Solving a longstanding problem on equiangular lines, we determine, for each given fixed angle and in all sufficiently large dimensions, the maximum number of lines pairwise separated by the given angle. A key ingredient is a new result in spectral graph theory: the adjacency matrix of a connected bounded degree graph has sublinear second eigenvalue multiplicity. My talk will discuss these problems and their connections. Here is an open problem that I would like to understand better: what is the maximum possible second eigenvalue multiplicity of a connected bounded degree graph?

Ginestra Bianconi – *The dynamics of higher-order networks: the effect of topology and triadic interactions*

Higher-order networks capture the interactions among two or more nodes and they are ubiquitous in complex systems. Here we show that higher-order interactions are responsible for new dynamical processes that cannot be observed in pairwise networks. We will cover how topology is key to define synchronization of topological signals, i.e. dynamical signal defined not only on nodes but also on links, triangles and higher-dimensional simplicies in simplicial complexes. Moreover we will reveal how triadic interactions can turn percolation into a fully-fledged dynamical process in which nodes can turn on and off intermittently in a periodic fashion or even chaotically leading to period doubling and a route to chaos of the percolation order parameter.

Filippo Radicchi – *Identifying the optimal embedding dimension of networks*

Network embedding is a general-purpose machine learning technique that encodes network structure in vector spaces with tunable dimension. Choosing an appropriate embedding dimension – small enough to be efficient and large enough to be effective – is challenging but necessary to generate embeddings applicable to a multitude of tasks. In this talk, I introduce a principled method for the selection of an optimal value of the embedding dimension such that all structural information of a network is parsimoniously encoded. The method is validated on various embedding algorithms and a large corpus of real-world networks. The embedding dimension selected by the introduced method in real-world networks suggest that efficient encoding in low-dimensional spaces is usually possible.

Peter Orbanz – *Limit theorems for distributions invariant under a group of transformations*

Consider a large random structure – a random graph, a stochastic process on the line, a random field on the grid – and a function that depends only on a small part of the structure. Now use a family of transformations to 'move' the domain of the function over the structure, collect each function value, and average. I will present results that show that, under suitable conditions, such transformation averages satisfy a law of large numbers, a central limit theorem, and a Berry-Esseen type bound on the speed of convergence. These imply results graphon models of networks, stationary random fields, and so forth. One relevant condition is that the distribution of the random structure remains invariant under the transformations, which can be read as a probabilistic symmetry property. Loosely speaking: The large-sample theory of i.i.d. averages still holds if the i.i.d. assumption is relaxed to a symmetry assumption.

François Caron – *Sparse graphs and multigraphs based on exchangeable random measures*

Random simple and multigraph models based on exchangeable random measures, also called (multi)graphex processes or generalised graphon models, have recently been proposed as a flexible class of sparse random graph models. I will present this class of models and discuss some of their asymptotic properties, in particular the asymptotic behaviour of the degree distribution.

Diego Garlaschelli – *Multiscale network renormalization: scale-invariance without geometry*

Systems with lattice geometry can be renormalized exploiting their coordinates in metric space, which naturally define the coarse-grained nodes. By contrast, complex networks defy the usual techniques, due to their small-world character and lack of explicit geometric embedding. Current network renormalization approaches require strong assumptions (e.g. community structure, hyperbolicity, scale-free topology), thus remaining incompatible with generic graphs and ordinary lattices. Here we introduce a graph renormalization scheme valid for any hierarchy of coarse-grainings, thereby allowing for the definition of 'block-nodes' across multiple scales. This approach

reveals a necessary and specific dependence of network topology on additive hidden variables attached to nodes, plus optional dyadic factors. Renormalizable networks turn out to be consistent with a unique specification of the fitness model, while they are incompatible with preferential attachment, the configuration model or the stochastic blockmodel. These results highlight a deep conceptual distinction between scale-free and scale-invariant networks, and provide a geometry-free route to renormalization. If the hidden variables are annealed, they lead to realistic scale-free networks with density-dependent cut- off, assortativity and finite local clustering, even in the sparse regime and in absence of geometry. If they are quenched, they can guide the renormalization of real-world networks with node attributes and distance-dependence or communities. As an application, we derive an accurate multiscale model of the International Trade Network applicable across hierarchically nested geographic partitions.

Giovanni Petri – *Between higher-order mechanisms and phenomena*

Complex networks have become the main paradigm for modelling the dynamics of complex interacting systems. However, networks are intrinsically limited to describing pairwise interactions, whereas real-world systems are often characterized by higher-order interactions involving groups of three or more units. Higher-order structures, such as hypergraphs and simplicial complexes, are therefore a better tool to map the real organization of many social, biological and man-made systems. At the same time, higher-order observables, typically topological or information-theoretic in nature and often sharing the same simplicial language, have been gathering attention for their capacity to capture properties of complex systems that are invisible to standard statistical descriptions. This had led to a certain confusion between these two facets, mechanisms on one side, phenomena on the other. Here, using recent examples from both computational modeling and neuroimaging analysis, I highlight collective behaviours induced by higher-order interactions, their interface with recent advances in topological data analysis, and finally outline three key challenges for the physics of higher-order complex systems.

Marta Sales-Pardo – *Using inference for recommendation and to obtain models from data: Is data always good enough?*

I will talk about two different problems that share a common behavior: a transition between a desired inference outcome and a less desirable one. My first example example will be that of recommender systems (or bipartite graphs with multi-valued edges) and the use of node attributes (such as gender of a user and the genre of a movie) to increase prediction accuracy of unobserved ratings. My second example will be that of obtaining models from data using a Bayesian inference framework in the presence of noise.

I will discuss how data attributes are not always useful to make recommendations in the same way that increasing the noise in the data will prevent us from finding the model that generated the data. I will also discuss that if we increase the importance of the attributes or the noise, we observe a transition between a regime in which we only see the data and a regime in which we only see the attributes or models that are compatible with noise.

Clara Stegehuis – *Optimal structures in complex networks: where can we find them?*

Subgraphs contain important information about network structures and their functions. But where in the network can we find them? We investigate this for several random graph models, by using an optimization problem that identifies the dominant structure of any given subgraph. The optimizer describes the degrees of the vertices that together create the most likely subgraph. On the popular hyperbolic random graph model, our optimization problem shows the trade-off between geometry and popularity: some subgraphs are most likely formed by vertices that are close by, whereas others are most likely formed by vertices of high degree. We then show that this optimization problem easily extends to other network structures, such as local clustering, which expresses the probability that two neighbors of a vertex are connected.

Tiago Peixoto – *Disentangling homophily, community structure and triadic closure*

One of the most typical properties of social networks is the presence of homophily, i.e. the increased tendency of an edge to exist between two nodes if they share the same underlying characteristic, such as race, gender, class and a variety of other social parameters. More broadly, when the underlying similarity parameter is not specified a priori, the same homophily pattern is known as community structure. Another pervasive pattern encountered in the same kinds of network is transitivity, i.e. the increased probability of observing an edge between two nodes if they have a neighbor in common. Although these patterns are indicative of two distinct mechanisms of network formation, namely choice or constraint homophily and triadic closure, respectively, they are generically conflated in non-longitudinal data. This is because both processes can result in the same kinds of observation: 1. the preferred connection between nodes of the same kind can induce the presence of triangles involving similar nodes, and 2. the tendency of triangles to be formed can induce the formation of groups of nodes with a higher density of connections

between them, when compared to the rest of the network. This conflation means we cannot reliably interpret the underlying mechanisms of network formation merely from the abundance of triangles or observed community structure in network data.

In this talk I present a solution to this problem, consisting in a principled method to disentangle homophily and community structure from triadic closure in network data. This is achieved by formulating a generative model that includes community structure in a first instance, and an iterated process of triadic closure in a second. Based on this model, we develop a Bayesian inference algorithm that is capable of identifying which edges are more likely to be due to community structure or triadic closure, in addition to the underlying community structure itself. As we show, this reconstruction yields a detailed interpretation of the underlying mechanisms of network formation, allowing us to identify macro-scale structures that emerge spontaneously from micro-scale higher-order interactions, and in this way we can separate them from inherently macro-scale structures. We show how the method can evade the detection of spurious communities caused solely by the formation of triangles in the network, and how it can improve the performance of link prediction when compared to the pure version of the SBM without triadic closure.

Thomas Bläsius – *Theoretical algorithm analysis meets practical data*

A theoretical running time analysis is a great tool for predicting the performance of algorithms, saving programmers from implementing and evaluating alternatives that are "obviously" worse. Though this works very well in many cases, there are situations where the predictions do not match practical observations. This often comes from the fact that practical problem instances have certain properties that make them well-behaved. In this talk, I discuss how this can be formalized using probabilistic network models and whether the resulting theoretical predictions coincide with observations on practical data.

Marián Boguñá – *Topological phase transitions in the geometric configuration model*

The (soft) configuration model (CM) has been extremely successful as a null model for real networks. Given a degree sequence from a real network, the CM is defined as the maximally random graph ensemble with that given (expected) degree sequence. A remarkable property of this model is the fact that interactions among nodes are pairwise. In its soft version, this is equivalent to say that any pair of nodes i, j are connected independently with probability $p_{ij} \sim N^{-1}\kappa_i\kappa_j$, with κ_i and κ_j accounting for the expected degrees of nodes i and j. However, the CM model is unable to generate finite clustering in the thermodynamic limit because the connection probability is inversely proportional to the system size.

To overcome this problem, we introduced the network geometry paradigm, which main hypothesis states that the architecture of real complex networks has a geometric origin. The nodes of the complex network can be characterized by their positions in an underlying metric space so that the observable network topology–abstracting their patterns of interactions–is then a reflection of distances in this space. This simple idea led to the development of a very general framework able to explain the most ubiquitous topological properties of real networks, namely, degree heterogeneity, the small-world property, and high levels of clustering. Network geometry is also able to explain in a very natural way other non-trivial properties, like self-similarity and community structure, their navigability properties, and is the basis for the definition of a renormalization group in complex networks. Quite strikingly, these results are achieved with only pairwise interactions among nodes, while higher order structures are naturally induced by the underlying metric space.

Within this paradigm, the (soft) geometric configuration model (GCM) is defined as the maximally random ensemble of geometric random graphs able to generate graphs with a given (expected) degree sequence that are simultaneously sparse, small-world, clustered, and without degree-degree correlations. Clustering in the GCM undergoes a phase transition between a geometric phase with finite clustering coefficient in the thermodynamic limit and a non-geometric phase where the clustering coefficient is zero. This transition, however, does not fit within the Landau symmetry breaking paradigm. Instead, it is a topological phase transition between two different topological orders. Upon mapping the network ensemble to a system of noninteracting fermions at temperature β^{-1}, we find an anomalous behavior for the entropy of the ensemble, which diverges at the critical point. This leads to an anomalous scaling behavior for finite systems and to the definition of an effective system size scaling logarithmically with the number of nodes.

Maksim Kitsak – *Finding shortest and nearly shortest path nodes in large substantially incomplete networks*

Dynamic processes on networks, be it information transfer in the Internet, contagious spreading in a social network, or neural signaling, take place along shortest or nearly shortest paths. Unfortunately, our maps of most large networks are substantially incomplete due to either the highly dynamic nature of networks, or high cost of network measurements, or both, rendering traditional path finding methods inefficient. We find that shortest paths

in large real networks, such as the network of protein-protein interactions (PPI) and the Internet at the autonomous system (AS) level, are not random but are organized according to latent-geometric rules. If nodes of these networks are mapped to points in latent hyperbolic spaces, shortest paths in them align along geodesic curves connecting endpoint nodes. We find that this alignment is sufficiently strong to allow for the identification of shortest path nodes even in the case of substantially incomplete networks. We demonstrate the utility of latent geometric path finding in problems of cellular pathway reconstruction and communication security.

Fragkiskos Papadopoulos – *Dynamics of cold versus hot random hyperbolic graphs*

Random hyperbolic graphs (RHGs) are adequate models for real-world complex networks, as they naturally and simultaneously possess many of their common structural properties. However, most existing work on RHGs is focused on structural properties of network snapshots, i.e., of static graphs, while little is known about the dynamical properties of RHGs. In this talk, we will derive the most basic properties of dynamic RHGs, starting from the cold regime (network temperature in the model $T < 1$). We will show that the distributions of contact and intercontact durations decay as power laws with exponents $2 + T$ and $2 - T$, respectively. These results are qualitatively consistent with what observed in real-world temporal networks, such as human proximity networks. Interestingly, these results hold irrespective of the nodes' expected degree distribution, suggesting that the emergence of broad (inter)contact distributions in real systems is due to node similarities. We will also illustrate that many other properties of dynamic cold RHGs (e.g., group size distributions, abundance of recurrent components, etc.) are also consistent with real-world systems, justifying why epidemic and rumour spreading processes perform remarkably similar in real and modelled networks. We will also discuss recent results on dynamic RHGs in the hot regime ($T > 1$). Here, the configuration model emerges as an infinitive temperature limit. We will show that in this regime (inter)contact distributions can emerge purely due to the heterogeneity of expected node degrees. However, the intercontact distributions are always nonnormalizable, suggesting that hot RHGs (including the configuration model) cannot be used as null temporal network models. We will conclude with open research questions.

Johannes Lengler – *Information spreading in geometric inhomogeneous random graphs*

I will introduce our latest work on Geometric Inhomogeneous Random Graphs (GIRGs). The GIRG model combines a scale-free degree distribution with a spatial component, similar to the related models of Hyperbolic Random Graphs and Scale-Free Percolation. We have studied the following spreading process on this model: initially, only a single vertex is informed/infected, and the information is transmitted along the edges. The transmission time along an each edge is the product of two factors: a random component that is drawn i.i.d. for each edge; and a penalty factor that slows down transmission to and from vertices of large degree. Other than for classical spreading processes, we obtain an extremely rich phase diagram, with at least three phase transitions as the penalty term increases. The time to infect a vertex in distance d of the source can either be constant ("explosion"), or it can grow poly-logarithmically in d, or polynomially with an exponent less than one, or linear. Yet more regimes can appear inside of the phase transitions. I will argue that such a rich set of behaviors is in line with the large variety of spreading processes in real-world networks.

Miklós Rácz – *Correlated stochastic block models: graph matching and community recovery*

I will discuss statistical inference problems on edge-correlated stochastic block models. We determine the information-theoretic threshold for exact recovery of the latent vertex correspondence between two correlated block models, a task known as graph matching. As an application, we show how one can exactly recover the latent communities using multiple correlated graphs in parameter regimes where it is information-theoretically impossible to do so using just a single graph. Furthermore, we obtain the precise threshold for exact community recovery using multiple correlated graphs, which captures the interplay between the community recovery and graph matching tasks. This is based on joint work with Julia Gaudio and Anirudh Sridhar.

Subhabrata Sen – *Incorporating network side information in statistical learning: a statistical physics based approach*

Network side-information is often available in modern supervised and unsupervised problems. Incorporating this side-information with other data sources is an interesting challenge in modern data science. In this talk, we will discuss a principled approach to this end, using some ideas arising in statistical physics/graphical models. First, we discuss clustering in the 'contextual block model', where one wishes to cluster high-dimensional datapoints, given auxiliary pairwise measurements in the form of a network. This is based on joint work with Yash Deshpande, Andrea Montanari, Elchanan Mossel and Chen Lu. Time permitting, we will also discuss ongoing work on a supervised learning problem with network side information. This is based on joint work with Sagnik Nandy.

Souvik Dhara – *Spectral recovery of binary censored block models*

We consider a model of censored community detection, where most of the data about the network is missing as the status of only a small fraction of the potential edges is revealed. Our main interest is to understand the performance of spectral algorithms based on top eigenvectors of a weighted, signed adjacency matrix. The spectral algorithm is shown to succeed in recovering the communities exactly up to the information theoretic threshold for the symmetric Stochastic Block Model and the Planted Dense Subgraph problem. However, in other parameter regimes, we show that the spectral algorithm may not succeed up to the information theoretic threshold.

Paul Krapivsky – *Dynamic space packing*

The sphere packing problem asks for the densest packing. Other packing problems arise in coding theory, material science, etc. Little is known about the properties of the packings, e.g., about the underlying network of touching spheres, particularly in high dimensions. I will describe a very simple packing process in which space is covered by balls that are added at a constant rate. Balls overlapping with an added ball are immediately removed. In a lattice version, sites are filled independently and uniformly, and after every deposition event each neighboring site is emptied. The steady-state occupancy and other basic statistical features, e.g., the pair correlation functions, can be analytically computed on $\mathbb{R}^d$ and $\mathbb{Z}^d$ in arbitrary dimension.