

Properties of Random Graphs with Hidden Color

Bo Söderberg*

Complex Systems Division, Dept. of Theoretical Physics, Lund University, Sweden

(Dated: 9th May 2003)

We investigate in some detail a recently suggested general class of ensembles of sparse undirected random graphs based on a hidden stub-coloring, with or without the restriction to nondegenerate graphs. The calculability of local and global structural properties of graphs from the resulting ensembles is demonstrated. Cluster size statistics are derived with generating function techniques, yielding a well-defined percolation threshold. Explicit rules are derived for the enumeration of small subgraphs. Duality and redundancy is discussed, and subclasses corresponding to commonly studied models are identified.

PACS numbers: 02.50.-r, 64.60.-i, 89.75.Fb

Keywords: random graph; network; statistical ensemble; phase transition; critical phenomena; percolation threshold; subgraph count

I. INTRODUCTION

Numerous phenomena in physics, molecular biology, social sciences and information technology can be described in terms of networks, where the nodes represent elementary units such as spins, genes, proteins, people, or computers, while the links describe their interaction structure. The formation process of these networks typically is not entirely deterministic, but involves stochastic components, and the resulting networks can be viewed as *random graphs* – random members of a *statistical ensemble* of graphs.

We are primarily interested in *truly random* graphs, without any prior distinction between individual nodes or groups of nodes, such as an underlying lattice or other regular structure. An example is the *classic model* of Erdős and Rényi [1], with a single parameter (in addition to the order N of the graph) in the form of a real number c , such that each possible edge is independently and randomly realized with a probability $p = c/N$ (in the sparse version). The classic model has been thoroughly studied in various versions, static as well as evolving [2–4]. Its asymptotic ($N \rightarrow \infty$) degree distribution is Poissonian with average c , and it displays a *phase transition* in the form of a percolation threshold at $c = 1$, above which a giant component emerges containing a finite fraction of the nodes in the thermodynamic limit of large N . For a long time this and related models dominated the stage; however, they fail to describe the properties of most real-world networks.

In the last decades a multitude of alternative random graph models have been investigated, falling into two major categories. In a *static model* a statistical ensemble of random graphs is considered with-

out bothering about how the graphs were formed [5–10]. A *dynamical model* attempts to describe the random growth and evolution of a network, leading to an evolving ensemble of graphs [11–14].

Here we will focus on *static descriptions* of random graphs in terms of fixed statistical ensembles, bearing in mind that the dynamics of real-world networks is not always directly observable, and the comparison of model and reality typically has to be done based on static properties as observed in snapshots of real networks.

For the inference of a particular model based on the observed properties of real networks to be meaningful, a sufficiently general formalism is desirable, where more specific models appear as special cases of one and the same general class of graph ensembles.

In a recent paper [15], a promising candidate for such a general formalism was proposed; it will be referred to as *CDRG* (for Colored Degree-based Random Graphs). It is based on a *hidden coloring* of *stubs* (incidence points of edges upon a vertex) and a specification of the colored stub distribution of vertices as well as edges. This approach admits a unifying formalism for models of symmetric, truly random graphs that are *sparse* (typical degrees are finite and do not grow with the graph size N).

The resulting class of random graph ensembles incorporates several commonly studied models, such as the classic random graph, and random graphs with a given degree distribution [5, 7, 16, 17], as well as vertex-colored extensions of these [8, 10]. Models with degree-biased edge distributions [9] also fit into this approach. Furthermore, although the approach in its present form is restricted to symmetric graphs, it has a natural extension to directed graphs, which will be explored in forthcoming work.

The discussion in ref. [15] was restricted to ensembles of *simple* (nondegenerate) graphs containing no cycles of length one (self-couplings, or tadpoles) or two (double edges), based on the restriction to sim-

*Electronic address: Bo.Soderberg@thep.lu.se

ple graphs of an underlying ensemble of *multigraphs* where such degeneracies are allowed. Multigraph ensembles are interesting in their own right, and more convenient for analytical purposes. Here, we will consider both types of CDRG ensembles, denoting by *CDRG-s* the restriction to the class of ensembles of simple graphs, and by *CDRG-m* the unrestricted class of multigraph ensembles. For generic ensembles of both types, we will present a theoretical analysis of the properties of the resulting graphs, with an emphasis on the analysis of observable local and global graph characteristics.

The computability of structural properties is an important factor for the possibility of devising a systematic model inference scheme based on the observed properties of real-world networks. Both types of ensemble admit an analysis of both global and local structural properties of the resulting random graphs. The global connectivity properties of a graph can be analyzed in terms of the size distribution of connected components, for which a generating function analysis was devised in ref. [15]. Local structural properties are associated with the frequencies of appearance of small subgraphs; also these will be shown to be asymptotically computable in both types of ensemble.

The remainder of this article has the following structure. In section II, we will define our notation and introduce basic concepts to be used in the rest of the paper. Questions regarding ensemble definitions, for CDRG-m as well as CDRG-s, will be discussed in section III. Section IV contains a basic statistical analysis of the ensembles as seen from the point of view of the stubs. In section V, we will discuss the statistics of the number of copies of an arbitrary small graph as a subgraph of a random graph, and define rules for the computation of the asymptotically expected counts, pointing out differences and similarities between CDRG-s and CDRG-m ensembles. In section VI, we will discuss the global properties of random graphs from CDRG ensembles, as revealed by a generating function analysis of the cluster size distribution, extending the analysis presented in ref. [15]. Both the global and local analysis reveal a certain redundancy (symmetry) property of CDRG models, which forms the subject of section VII. In section VIII we will identify subclasses of CDRG ensembles corresponding to commonly studied models. Section IX, finally, contains a resumé of our main results, and some concluding remarks and speculations.

II. NOTATION AND BASIC CONCEPTS

A *labelled graph* consists of a set of distinguishable *vertices* (nodes, sites, points), which may be pairwise

connected by *edges* (links, bonds, lines).

Unless otherwise stated, a graph is assumed to be *symmetric* (undirected), such that edges have no particular direction (as opposed to a *digraph* – or directed graph – where an edge has a direction, pointing from one vertex to another).

A graph with N vertices is conveniently represented by its symmetric $N \times N$ *adjacency matrix* \mathbf{S} . An element S_{ij} counts the number of edges between vertices i and j ; thus, each edge contributes both to S_{ij} and S_{ji} ; as a result each diagonal element S_{ii} will be even, representing twice the number of self-couplings of vertex i . In a *simple* graph, cycles of length one (self-couplings or tadpoles) and two (multiple edges) are absent; as a result, the diagonal elements of \mathbf{S} are zero, and the remaining elements are restricted to the values zero or one. A *multigraph* may be simple or degenerate.

The *degree* (or connectivity) m of a vertex is defined as the number of edges connected to it, given by the corresponding row sum $\sum_j S_{ij}$; the vertex can be considered as possessing m *stubs* – points where a single edge endpoint (*butt*) is attached.

It is sometimes convenient to consider not only the vertices, but also the edges, and indeed the individual stubs and butts, as being distinguishable.

The *degree sequence* of a graph is an ordered list of N integers $(m_1 \dots m_N)$, describing the individual degrees of the N vertices. Alternatively, it can be summarized in terms of the *degree counts*, $N_m = \sum_i \delta(m, m_i)$, counting the number of vertices having degree m .

A commonly studied class of ensembles is based on giving an asymptotic *degree distribution* $\{p_m\}$, from which a compatible degree sequence can be determined for a given graph size N with degree counts $N_m \approx Np_m$. Then a random compatible graph is chosen by means of a random stub pairing (the *configuration model* [5, 17]). This approach will be referred to as *DRG*, for degree-based random graphs.

In another approach, *IRG* (for inhomogeneous random graphs), a class of vertex-colored extensions of the classic model has been considered, where each vertex is randomly and independently assigned an abstract type (*color*) drawn from a given distribution, and where edge probabilities are allowed to depend on the connected pair of colors [8].

In a recent article [15], the philosophies behind DRG and IRG were combined in a novel approach, where a hidden *stub-coloring* was used to define a very general class of ensembles with a given degree distribution. This approach, CDRG, forms the main subject of this article.

Thus, we will consider *stub-colored* graphs, where each stub independently carries an internal characteristic, a *hidden color* $a \in [1, \dots, K]$, to be considered unobservable. The degree m of a vertex

then decomposes into the sum of contributions m_a counting the stubs with a definite color a . These sub-degrees can be collected in a K -vector $\mathbf{m} = (m_1 \dots m_K)$, to be referred to as the *colored degree* of the vertex.

It is then natural to consider the *colored degree sequence* of such a graph, in terms of the numbers $N_{\mathbf{m}}$ of vertices with a distinct colored degree \mathbf{m} .

Accordingly, each edge connects a pair of colored stubs and can be associated with a color pair (a, b) . We can then also consider the count $n_{ab} = n_{ba}$ of edges for each color pair, where an ab -edge for practical reasons contributes both to n_{ab} and n_{ba} (so diagonal elements n_{aa} are even).

The total number of stubs with color a in the graph is then given by $\sum_b n_{ab}$; this must match the corresponding stub count $M_a \equiv \sum_{\mathbf{m}} m_a N_{\mathbf{m}}$. In particular, the total butt count, $\sum_{ab} n_{ab}$, must be even (being twice the number of edges), and it must equal the total stub count, $M = \sum_a M_a = \sum_{\mathbf{m}} \sum_a m_a N_{\mathbf{m}}$. We will find it convenient to collect the colored stub counts in a vector $\mathbf{M} = (M_1, \dots, M_K)$.

Throughout this article, K -vectors will be denoted by (mostly lower case, with \mathbf{M} being an exception) fat symbols such as $\mathbf{x} \equiv (x_1 \dots x_K)$, in terms of which an obvious simplified notation will be used: $\mathbf{x}^{\mathbf{m}} \equiv \prod_a x_a^{m_a}$, $\mathbf{m}! \equiv \prod_a m_a!$, etc. The uniform K -vector $(1, \dots, 1)$ will be denoted as $\mathbf{1}$. Similarly, $K \times K$ -matrices will be denoted by upper case fat symbols such as $\mathbf{T} = \{T_{ab}\}$, with matrix product indicated by juxtaposition. A *component-wise* product will be denoted by a cross (\times), as in $\mathbf{x} \times \mathbf{m} \equiv (x_1 m_1, \dots, x_K m_K)$. The transpose of a matrix \mathbf{T} will be denoted by \mathbf{T}^{\top} and the matrix inverse of the transpose by $-\mathbf{T}^{\top}$.

We will be interested in models based on a definite *colored degree distribution* (CDD) $\{p_{\mathbf{m}}\}$, in terms of which we can define moments $\langle m_a \rangle = \sum_{\mathbf{m}} p_{\mathbf{m}} m_a$, etc. Such a distribution is conveniently described by its multivariate *generating function*,

$$H(\mathbf{x}) = \sum_{\mathbf{m}} p_{\mathbf{m}} \mathbf{x}^{\mathbf{m}}, \quad (1)$$

where $\mathbf{x} = (x_1 \dots x_K)$ is a K -component vector of auxiliary variables.

From H the individual $p_{\mathbf{m}}$ can be extracted by means of repeated differentiation at $\mathbf{x} = \mathbf{0}$, while repeated differentiation at $\mathbf{x} = \mathbf{1}$ yields the *combinatorial moments*

$$E_{ab\dots} = \partial_a \partial_b \dots H(\mathbf{x} = \mathbf{1}), \quad (2)$$

where ∂_a stands for the derivative with respect to x_a . Thus, the lowest moments become $E_a = \langle m_a \rangle$, $E_{ab} = \langle m_a m_b - m_a \delta_{ab} \rangle$, etc., generalizing the corresponding combinatorial moments of the *total* degree, $\langle m \rangle$, $\langle m(m-1) \rangle$, etc. Occasionally we will

suppress indices and refer to the n th order combinatorial moment as $\mathbf{E}_{(n)}$. Thus, $\mathbf{E}_{(1)} = \{E_a\} = \langle \mathbf{m} \rangle$, $\mathbf{E}_{(2)} = \{E_{ab}\}$, $\mathbf{E}_{(3)} = \{E_{abc}\}$, etc. In particular, it is frequently convenient to view the second order tensor $\mathbf{E}_{(2)}$ as a *matrix*, denoted simply by \mathbf{E} .

Upon summing over the indices independently, the n th order *scalar* combinatorial moments result, denoted $E_{(n)}$. Thus, $E_{(1)} = \sum_a E_a = \sum_a \langle m_a \rangle = \langle m \rangle$, $E_{(2)} = \sum_{ab} E_{ab} = \langle m(m-1) \rangle$, $E_{(3)} = \sum_{abc} E_{abc} = \langle m(m-1)(m-2) \rangle$, etc.

III. MODEL DEFINITIONS

Ensembles in CDRG are based on *asymptotic models*, where a desired asymptotic behaviour as $N \rightarrow \infty$ is specified. For a given asymptotic model, finite graph ensembles can be defined.

A. Asymptotic CDRG model

An asymptotic model is defined as follows.

Asymptotic CDRG model:

- Specify the desired color space, taken to be $[1, \dots, K]$ for some integer $K \geq 1$;
- Choose a normalized asymptotic colored degree distribution $\{p_{\mathbf{m}}\}$, with $p_{\mathbf{m}} \geq 0$ and $\sum_{\mathbf{m}} p_{\mathbf{m}} = 1$;
- Choose a symmetric $K \times K$ *color preference matrix* \mathbf{T} , with real, non-negative elements $T_{ab} \geq 0$, subject to the constraint

$$\sum_b T_{ab} \langle m_b \rangle = 1. \quad (3)$$

The role of \mathbf{T} is to control the asymptotic symmetrized color-specific distribution of edges: $n_{ab} \sim N \langle m_a \rangle T_{ab} \langle m_b \rangle$, where n_{ab} denotes the number of edges connecting colors a and b . The constraint (3) is needed for the mutual consistency between the asymptotic vertex and edge statistics – roughly speaking, it secures a matching butt for each stub.

Following ref. [15], we will for simplicity assume colored degree distributions to be well-behaved, such that all moments of arbitrary order are defined. This excludes power tails in the degree distribution – the particular complications associated with extending CDRG to *fat-tailed* distributions fall outside the scope of this article, and will hopefully be the subject of a future paper.

B. Ensembles of finite graphs

Based on a given asymptotic model, we wish to define an ensemble of multigraphs or simple graphs with a given size N .

1. Multigraph ensembles – CDRG- m

The simplest and most straightforward way to define an ensemble of *multigraphs* of a given size N consistently with a given asymptotic CDRG model is as follows. Fix the color-specific vertex and edge counts, $N_{\mathbf{m}}$ and n_{ab} , as close as possible to their expected values, i.e. $N_{\mathbf{m}} \approx N p_{\mathbf{m}}$, and $n_{ab} \approx N \langle m_a \rangle T_{ab} \langle m_b \rangle$, such that they yield matching colored stub and butt counts, $\sum_{\mathbf{m}} N_{\mathbf{m}} m_a = \sum_b n_{ab} = M_a \approx N \langle m_a \rangle$. Then place edges by for each color a randomly pairing each of the M_a stubs with a unique matching butt.

The result can be considered a *microcanonical ensemble* of multigraphs, and was used in the original article [15] as a means to define an ensemble of simple graphs by projecting out the simple part.

In this article, we will consider a slightly different multigraph ensemble where only N is fixed while the other counts are allowed to vary. While being slightly more elaborate to implement as a random graph generator, this *grand canonical* ensemble is more convenient for analytical purposes.

Grand canonical multigraph ensemble

1. For each of the N vertices, draw its colored degree at random from the asymptotic distribution $\{p_{\mathbf{m}}\}$. The result is a random colored degree sequence, yielding a definite stub count M , the expected value of which is $N \langle m \rangle$. Repeat this step until M is even.
2. Consider the entire set of $(M - 1)!!$ pairings of the M stubs, and associate with each pairing a statistical weight given by the product of single edge factors, where each ab -edge contributes a factor T_{ab}/N . Draw a pairing at random from the resulting weighted distribution.

The weighted random pairing defines a natural colored extension of the stub-pairing method, the configuration model, as used in DRG [17].

In the thermodynamic limit, the microcanonical and grand canonical ensembles corresponding to the same asymptotic model should be statistically equivalent. Indeed, when $N \rightarrow \infty$, the distribution of colored degree counts $N_{\mathbf{m}}$ in the grand canon-

ical ensemble becomes sharply peaked around the microcanonical values $\langle N_{\mathbf{m}} \rangle = N p_{\mathbf{m}}$. As a result, the total colored stub counts M_a will be close to $N \langle m_a \rangle$, and as will be shown below, this implies that the distribution of colored edge counts n_{ab} resulting from the weighted pairing becomes sharply peaked around the microcanonical ensemble values $\langle n_{ab} \rangle = N \langle m_a \rangle T_{ab} \langle m_b \rangle$. In the next section we will give a detailed analysis of the basic stub pairing statistics.

2. Simple graph ensembles – CDRG- s

In ref. [15], a microcanonical ensemble of simple graphs was defined by projecting out the simple graph part from the microcanonical ensemble of multigraphs, as realized by redoing the random butt-stub pairing step until a simple graphs results.

Here, we shall instead consider a *grand canonical ensemble* of simple graphs, defined by projecting out the simple part from the corresponding CDRG- m ensemble. It can be realized e.g. by repeatedly drawing a member of the latter until a nondegenerate graph results.

The efficiency of this method depends on the probability for a randomly drawn multigraph to be simple. This probability is easily computed, as will be demonstrated below (in the section on local characteristics), where we will verify the result given in ref. [15].

In ref. [15] it was also argued that several statistical graph properties not directly involving the presence or absence of degeneracies as measured in a CDRG ensemble of simple graphs were asymptotically identical to those of the underlying multigraph ensemble; we shall provide arguments that this is indeed the case.

IV. BASIC STUB STATISTICS

For the forthcoming analysis of local and global structural properties of random graphs drawn from the grand canonical ensemble of multigraphs, an initial basic statistical analysis of the graph properties as seen from the point of view of the individual stubs is required.

A. Colored stub distribution

In a grand canonical CDRG- m ensemble, each vertex i can be considered to have an independent random colored degree \mathbf{m}_i drawn from the asymptotic distribution $\{p_{\mathbf{m}}\}$ (neglecting the slight modification due to the constraint of even M). Hence, the

vector $\mathbf{M} = \sum_i \mathbf{m}_i$ of total colored stub counts is essentially the sum of N independent colored degrees, which trivially results in the \mathbf{M} distribution $P_{\mathbf{M}}$ being centered around the expected stub count $\langle \mathbf{M} \rangle = N \langle \mathbf{m} \rangle$, with fluctuations of $O(N^{1/2})$ governed by the correlation matrix $\langle \mathbf{M} \mathbf{M}^T \rangle_c = N \langle \mathbf{m} \mathbf{m}^T \rangle_c$.

For the derivation of more general properties of $P_{\mathbf{M}}$, it may be convenient to use its generating function, which is given by $H(\mathbf{z})^N = \sum_{\mathbf{M}} P_{\mathbf{M}} \mathbf{z}^{\mathbf{M}}$, [22] where $H(\mathbf{z})$ is the generating function for $p_{\mathbf{m}}$, as defined in eq. (1). From $H(\mathbf{z})^N$, $P_{\mathbf{M}}$ can be extracted as the coefficient for $\mathbf{z}^{\mathbf{M}}$:

$$P_{\mathbf{M}} = \oint \frac{d\mathbf{z}}{2\pi i \mathbf{z}} \mathbf{z}^{-\mathbf{M}} H(\mathbf{z})^N, \quad (4)$$

where $\oint \frac{d\mathbf{z}}{2\pi i \mathbf{z}}$ stands for $\prod_a \oint \frac{dz_a}{2\pi i z_a}$, denoting the complex integration of each z_a along a path encircling the origin. For \mathbf{M} close to its average $N \langle \mathbf{m} \rangle$, the integral is asymptotically dominated by the contributions from a saddlepoint $\mathbf{z} \approx \mathbf{1}$, from which the asymptotic properties of $P_{\mathbf{M}}$ can be derived in a saddlepoint approximation.

B. Stub pairing statistics

Next, we wish to analyze the result from the weighted random pairing of stubs. To that end we note that for a given assignment of colored vertex degrees, the only thing important for the pairing step is the total stub count $\mathbf{M} = \{M_a\} = \sum_i \mathbf{m}_i$.

Denote by $Z(\mathbf{M})$ the total weight of the set of $(M-1)!!$ possible stub pairings, given \mathbf{M} . It is the sum over distinct pairings π of the associated product of edge weights T_{ab}/N , and can be written as follows:

$$\begin{aligned} Z(\mathbf{M}) &= \sum_{\pi} \prod_{\text{pairs}} \frac{T_{ab}}{N} \\ &= N^{-M/2} \mathbf{M}! \sum_{\{n_{ab}\}} \prod_{a < b} \frac{T_{ab}^{n_{ab}}}{n_{ab}!} \prod_a \frac{T_{aa}^{n_{aa}/2}}{n_{aa}!!} \end{aligned} \quad (5a)$$

$$= N^{-M} \mathbf{M}! \oint \frac{d\mathbf{z}}{2\pi i \mathbf{z}} \mathbf{z}^{-\mathbf{M}} e^{\frac{N}{2} \mathbf{z}^T \mathbf{T} \mathbf{z}}, \quad (5b)$$

where the sum over $\{n_{ab}\}$ is restricted to non-negative, symmetric values with even diagonal and correct row sums, $\sum_b n_{ab} = M_a$. The last form, (5b), is obtained by Fourier-expanding the implicit Kronecker deltas for the row sum constraints.

So far, everything is exact. The complex integral form of $Z(\mathbf{M})$ can be estimated in a saddlepoint approximation, based on extremizing the associated ‘‘action’’, $S(\mathbf{z}) = \mathbf{M} \cdot \log(\mathbf{z}) - \frac{N}{2} \mathbf{z}^T \mathbf{T} \mathbf{z}$. Demanding a vanishing derivative, $\partial_{z_a} S = \frac{M_a}{z_a} - N \mathbf{T} \mathbf{z} = 0$, yields the equation for a saddlepoint as

$$\mathbf{M} = N \mathbf{z} \times (\mathbf{T} \mathbf{z}), \quad (6)$$

implicitly defining the saddlepoint $\mathbf{z}(\mathbf{M})$ (up to a total sign, really, but for even M , the two yield identical contributions).

For the particular choice of $\mathbf{M} = N \langle \mathbf{m} \rangle$, defining the expected value of \mathbf{M} , the relevant solution is $\mathbf{z} = \langle \mathbf{m} \rangle$, yielding for the total weight the asymptotic value $Z(\mathbf{M} = N \langle \mathbf{m} \rangle) \sim e^{-N \langle \mathbf{m} \rangle^2 / 2}$, where we have disregarded subexponential factors and assumed M to be even. The value of $Z(\mathbf{M})$ for slightly different arguments can then be estimated by noting that a small relative change in \mathbf{M} yields a small relative change in \mathbf{z} , and leads to a small change in the value of the action S .

Thus, upon replacing \mathbf{M} by a modified value $\hat{\mathbf{M}} = \mathbf{M} + \epsilon$, the saddlepoint \mathbf{z} changes to $\hat{\mathbf{z}} = \mathbf{z} + \delta$, and the action $S = S(\mathbf{M}, \mathbf{z})$ changes to $\hat{S} = S + \epsilon \cdot \partial S / \partial \mathbf{M} + \delta \cdot \partial S / \partial \mathbf{z}$, evaluated at $\mathbf{M} = N \langle \mathbf{m} \rangle$, $\mathbf{z} = \langle \mathbf{m} \rangle$, where the \mathbf{z} derivative vanishes due to the saddlepoint condition. Thus, to lowest order, the modified value of the action is given by $\hat{S} = S + \epsilon \cdot \log(\mathbf{z})$. As a result, the complex integral to leading order changes by a factor of $\mathbf{z}^{-\epsilon}$, and thus the total weight Z changes by a factor $(\mathbf{M}/N \mathbf{z})^{\epsilon} \approx 1$ – i.e. not at all. This means that $Z(\mathbf{M})$ has a saddlepoint for \mathbf{M} close to its expected value, $\langle \mathbf{M} \rangle = N \langle \mathbf{m} \rangle$.

C. Individual pairing probabilities

The asymptotic probability that an arbitrarily chosen pair of stubs will be connected in the random pairing, given their colors a, b , can be calculated as the ratio of the total weight *conditional* on this connection and the unconditional total weight. The conditional weight is obtained by multiplying the factor T_{ab}/N for the clamped edge by the total weight $Z(\mathbf{M} - \mathbf{e}_a - \mathbf{e}_b)$ of all pairings of the remaining $M - 2$ stubs, where \mathbf{e}_a denotes the unit vector along the positive a -direction. This is to be divided by $Z(\mathbf{M})$; as argued above the Z ratio is asymptotically 1, and so the asymptotic probability is simply T_{ab}/N .

Let us check this result for consistency: There are M_b stubs with color b ; each of these defines an equally probable matching partner to a fixed stub of color a (neglecting for the case $a = b$ the asymptotically negligible possibility that the two stubs be identical), yielding $T_{ab} M_b / N \approx T_{ab} \langle m_b \rangle$ for the probability that the pairing partner of an arbitrary stub of color a has color b . A final summation over b yields $\sum_b T_{ab} \langle m_b \rangle = 1$, expressing the correct normalization of the asymptotic probabilities.

The argument is easily extended to yield the asymptotic probability for an arbitrary finite number of clamped stub pairs in the grand canonical ensemble of multigraphs, as given simply by the product of the corresponding edge factors T_{ab}/N , with

the relative error being of order $O(N^{-1})$.

From these pairing probabilities we can draw the trivial conclusion that the colored edge counts n_{ab} in a grand canonical CDRG-m ensemble asymptotically will be close to the corresponding microcanonical ensemble values $N \langle m_a \rangle T_{ab} \langle m_b \rangle$; this can also be derived directly from eqs. (5).

Conversely, it is easily realized that asymptotically identical pairing probabilities hold for the microcanonical multigraph ensemble, where the colored edge counts are fixed to $n_{ab} \sim N \langle m_a \rangle T_{ab} \langle m_b \rangle$. Given an arbitrary pair of distinct stubs with respective colors a, b , the probability that they be paired is the product of (1) the probability n_{ab}/M_a that the first stub is chosen to belong to the group of a -stubs selected to be paired with color b , (2) the corresponding probability n_{ab}/M_b for the other stub, and (3), the probability $1/n_{ab}$ that the first stub is paired with the second among the n_{ab} candidates. Multiplying the three factors together yields the probability $n_{ab}/(M_a M_b) \sim T_{ab}/N$.

V. LOCAL CHARACTERISTICS

Calculability of local as well as global graph characteristics in a model greatly simplifies the task of model inference from observed graphs. All local graph characteristics can be derived from the embedding counts of various small connected *subgraphs*. These are easy to measure in observed graphs. The analysis given in the previous section provides the necessary tools for deriving rules for calculating the asymptotically expected count distributions in a CDRG model. We will first consider the case of a CDRG-m; the results for that case can then be used to derive the corresponding results for CDRG-s. Except where otherwise stated, the grand canonical ensembles will be assumed.

A. Subgraph statistics I: Multigraph ensemble

1. Initial discussion

Given an arbitrary, possibly degenerate, small graph γ with v vertices and e edges, we wish to study the statistics of the number n_γ of distinct copies of γ found in a random graph Γ drawn from a CDRG-m ensemble, i.e. the number of distinct subgraphs of Γ isomorphic to γ .

A subgraph of Γ is defined as a subset \mathbf{v} of the N vertices of Γ , together with a subset \mathbf{e} of the edges among \mathbf{v} . Two subgraphs are considered distinct if they have different \mathbf{v} or different \mathbf{e} . Note that a general subgraph is not necessarily an *induced* subgraph, where \mathbf{e} must be the entire set of edges among

\mathbf{v} . Thus, e.g., if γ lacks an edge between a pair of vertices, the corresponding pair in the target set \mathbf{v} may well be connected.

We are primarily interested in connected γ , but we will allow ourselves to consider also cases where γ is not connected. Let us begin by considering a few simple examples explicitly.

Single vertex (•): Let γ be a *single vertex with no edges*. Then we must have $n_\gamma = N$, since there are N ways to choose a single target vertex in Γ .

Unconnected pair of vertices (• •): Let γ consist of *two vertices and no edges* (so γ is not connected!). Then, $n_\gamma = N(N-1)/2 \sim N^2/2$, reflecting the $N(N-1)$ ways to choose an ordered pair of vertices in Γ , while the symmetry of γ under interchange of the two vertices makes the two *a priori* distinct orderings equivalent.

Connected pair (•—•): Let γ be the graph consisting of two vertices connected by a single edge. Again, γ is symmetric under interchange of its two vertices, and the target pair \mathbf{v} of vertices can be chosen in $N(N-1)/2$ distinct ways. Not all vertex pairs are connected, while others are multiply connected: A pair with k connections yields k distinct copies of γ . The average number of connections between an arbitrary pair of vertices is the sum over color pairs a, b of the average number of ab -edges connecting them. Each vertex of the pair has a colored degree randomly drawn from $\{p_{\mathbf{m}}\}$. For a given pair \mathbf{m}, \mathbf{m}' , there are $m_a m'_b$ possible ways to choose the a, b -edge, each yielding a probability T_{ab}/N . Averaging this over the colored degrees \mathbf{m}, \mathbf{m}' yields $\sum_{\mathbf{m}} p_{\mathbf{m}} \sum_{\mathbf{m}'} p_{\mathbf{m}'} m_a m'_b T_{ab}/N = \langle m_a \rangle \langle m_b \rangle T_{ab}/N$. Finally, summing over a, b gives $\langle \mathbf{m} \rangle^\top \mathbf{T} \langle \mathbf{m} \rangle / N = \langle m \rangle / N$ for the expected number of edges between a randomly chosen pair of vertices. Multiplying this by the number of ways to choose the pair of vertices yields $\frac{N}{2} \langle m \rangle$ for the asymptotically expected number of copies; this is precisely the expected number of edges, $\langle M \rangle / 2$, as it must be, since every edge defines a distinct copy of γ .

2. Expected count for general γ

For a more general graph γ , the expected count can be computed by multiplying the number of ways $\binom{N}{v}$ to choose the vertex target set \mathbf{v} by the expected number of copies using a fixed target set \mathbf{v} . The latter is obviously independent of \mathbf{v} when Γ is a random graph, and is the sum of the expected number of copies for each of the (naively $v!$) inequivalent orderings of \mathbf{v} , defined as the number of ways to choose the target set \mathbf{e} from the existing edges among \mathbf{v} (e.g., an ordered k -tuple of edges between a specific pair of vertices in \mathbf{v} can be chosen in $n!/(n-k)!$ distinct ways, if the target pair in \mathbf{v} is connected by

n edges).

In addition, if γ has a nontrivial isomorphism group (in terms of permutations of vertices as well as permutations and flips of edges), the result must be divided by a *symmetry factor* S_γ , given by the order of this group. It consists in two factors: One is given by the order of the *vertex* permutation symmetry of γ , the other by the order of the group of permutations and flips of *edges* with fixed vertices leaving γ invariant, yielding a factor of $n!$ for each pair of distinct vertices in γ connected by n edges, and a factor of $n!2^n = (2n)!!$ for each vertex with n tadpoles (requiring $2n$ stubs).

This results in the following “*Feynman*” rules for the calculation of the asymptotically expected number n_γ of copies of an arbitrary small graph γ in a large random graph Γ drawn from a CDRG- m ensemble:

Rules for calculating expected asymptotic subgraph counts $\langle n_\gamma \rangle$:

1. Label each stub in γ with an independent color index;
2. Associate with every vertex in γ with n stubs labelled a, b, \dots a factor given by N times the corresponding component $E_{ab\dots}$ of the n th order combinatorial moment $\mathbf{E}_{(n)}$;
3. Associate with each edge in γ a factor T_{ab}/N , where a, b are the color labels of the connected stubs;
4. Multiply together all vertex and edge factors, and sum the result over the stub colors.
5. Divide the result by the proper symmetry factor S_γ , to yield the expected count $\langle n_\gamma \rangle$.

Sketch of proof: The individual *vertex factor* decomposes into a factor of N for the number of ways to choose the target vertex in Γ , and a factor $E_{ab\dots}$, which takes some explaining. Consider a vertex in γ with two stubs, assigned colors a, b . The colored degree \mathbf{m} of the target vertex is drawn from $p_{\mathbf{m}}$, and the number of ways to pick two stubs with correct colors, given \mathbf{m} , is $m_a m_b$ if $a \neq b$, and $m_a(m_a - 1)$ if $a = b$. Averaging over \mathbf{m} yields E_{ab} . The result generalizes to an arbitrary number of stubs.

The *edge factor* T_{ab}/N represents the individual stub-stub connection probability as derived in the previous section; it ultimately stems from the corresponding factor in the weighted random pairing involved in the definition of the grand canonical ensemble.

The vertex part of the *symmetry factor* simply stems from the fact that the existence of a vertex permutation symmetry of γ implies a reduction of the naive number $N(N-1)\dots(N-v+1) \sim N^v$ of inequivalent choices of ordered target sets \mathbf{v} . Similarly, the edge part reflects the equivalence of naively distinct edge target sets \mathbf{e} for the same \mathbf{v} , differing only by the interchange of edges connecting the same pair of vertices, or by a flip of a single edge connecting a vertex to itself.

The same asymptotic rules can be derived for the case of the microcanonical multigraph ensemble using similar arguments.

In table I, the expected counts are given for subgraphs in the form of *chains*, *stars* and simple *cycles* of arbitrary length for a CDRG- m model, and for a plain DRG model (CDRG- m restricted to a single color) for comparison. Note the simplification occurring in the expression for the expected count for each *leaf* node with a single connection, due to the identity $\mathbf{T}(\mathbf{m}) = \mathbf{1}$: The vertex factor for the leaf and the single edge factor gives upon summation of the color label assigned to the single stub a factor $\sum_a N \langle m_a \rangle \times T_{ab}/N \equiv 1$, and their only effect is to increase the degree of the moment associated with the neighboring vertex by adding an index (b) that is simply summed over.

3. Scaling with N and edge correlations

Of obvious interest is how n_γ scales with N . The rules for the calculation of the expected count yield a factor of N for each vertex and a factor of N^{-1} for every edge, so the total power of N is $v - e$, which can also be expressed as *the number of mutually disconnected components in γ , minus the number of loops in γ* . For a connected γ , this yields 1 minus its number of loops. Thus if γ is a *tree*, the expected number of copies scales as $O(N)$, while for a *one-loop connected* γ the expected number scales as $O(1)$; for any connected γ with more than one loop there are asymptotically no copies at all, since the expected number is suppressed by factors of N .

Let us demonstrate with a few simple examples the increased correlation possibilities in CDRG models as opposed to a plain DRG model. First, we compare the counts for *triangles* (3-cycles, Δ), *wedges* (3-chains, Λ), and *edges* (2-chains, I). In a CDRG ensemble, their respective expected counts are

$$\langle n_\Delta \rangle = \frac{\text{Tr}(\mathbf{TE})^3}{6}, \quad (7a)$$

$$\langle n_\Lambda \rangle = \frac{NE}{2}, \quad (7b)$$

$$\langle n_I \rangle = \frac{N \langle m \rangle}{2}. \quad (7c)$$

Subgraph γ	k range	Vertices v	Edges e	Diff. $v - e$	Symm. factor S_γ	$\langle n_\gamma \rangle_{\text{CDRG}}$	$\langle n_\gamma \rangle_{\text{DRG}}$
k -star	$k \geq 2$	$k + 1$	k	1	$k!$	$NE_{(k)}/k!$	$NE_{(k)}/k!$
k -chain	$k \geq 2$	k	$k - 1$	1	2	$\frac{N}{2} \mathbf{1}^\top \mathbf{E}(\mathbf{TE})^{k-3} \mathbf{1}$	$\frac{N}{2} E (E/\langle m \rangle)^{k-3}$
k -cycle	$k \geq 3$	k	k	0	$2k$	$\text{Tr}(\mathbf{TE})^k / (2k)$	$(E/\langle m \rangle)^k / (2k)$

Table I: Asymptotically expected counts of subgraphs in the form of *stars*, *chains*, and simple *cycles* of arbitrary size as computed in a CDRG-m model and, for comparison, in a corresponding uncolored (DRG) model. The k -**star** consists of a single “hub” vertex connected to each of k leaf nodes by a single edge. The symmetry factor of $k!$ is due to permutations of the k leaves. The factors for the k leaves have been simplified as described in the text. For both CDRG and DRG, the resulting expected count can be written as $N \langle \binom{m}{k} \rangle$, as may have been expected - each vertex in Γ with $m \geq k$ stubs defines $\binom{m}{k}$ copies; in this case the expected count depends only on the plain degree distribution, $\{p_m\}$. The k -**chain** consists of k vertices connected into a chain by $k - 1$ edges. The symmetry factor of 2 is due to a flip of the entire chain. The two leaf factors for the endpoints have been simplified. As a result, the 4-chain is the first chain where the expected count shows a nontrivial dependence on \mathbf{T} , distinguishing CDRG from plain DRG for which the expected chain counts form a simple geometric series. The k -**cycle** consists of k vertices connected into a closed loop by k edges. The symmetry factor $2k$ is due to flipping ($\rightarrow 2$) and rotating ($\rightarrow k$) the vertex order in the cycle.

A *plain DRG* ensemble yields an identical expression for $\langle n_\Delta \rangle$ as well as for $\langle n_I \rangle$, while the triangle count becomes $\langle n_\Delta \rangle_{\text{DRG}} = E^3 / (6 \langle m \rangle^3)$, yielding the relation

$$\langle n_\Delta \rangle_{\text{DRG}} = \frac{\langle n_\Lambda \rangle^3}{6 \langle n_I \rangle^3}, \quad (8)$$

absent in a generic CDRG ensemble.

Similarly, the expected k -*chain* count is $\langle n_k \rangle = N \mathbf{1}^\top \mathbf{E}(\mathbf{TE})^{k-3} \mathbf{1} / 2$. In plain DRG, this simplifies to a geometric series, $NE^{k-2} / (2 \langle m \rangle^{k-3})$, which again can be expressed in terms of the wedge and edge counts:

$$\langle n_k \rangle_{\text{DRG}} = \langle n_\Lambda \rangle^{k-2} \langle n_I \rangle^{3-k}, \quad (9)$$

whereas in CDRG, this strict relation is absent.

A popular edge correlation measure in the literature is the so called *clustering coefficient* C , defined as the probability that two randomly chosen neighbors of a random vertex are connected [18, 19]. In not-so-sparse random graph models with an excessive amount of triangles, as can be anticipated to result with a power tail in the degree distribution, or in models based on an underlying regular structure, C can attain a finite value. This is not the case in the type of models we are considering here, and we expect C to decrease as $O(N^{-1})$. We can estimate C by comparing the expected counts for triangles (3-cycles) and 3-chains. Their ratio multiplied by 3 gives the estimate $C = \text{Tr}(\mathbf{TE})^3 / (NE)$. While this indeed scales as $O(N^{-1})$, the finite number NC has a nontrivial dependence on \mathbf{T} , allowing it to deviate from the DRG value of $E^2 / \langle m \rangle^3$.

These examples serve to illustrate the role of the hidden color in enabling a non-trivial edge correlation structure, and in lifting the simple relations between different subgraph counts present in DRG.

4. Beyond expected counts: Distribution shape

The expected count $\langle n_\gamma \rangle$ of a given subgraph γ gives only partial information on the count distribution. Of interest are also the actual shapes of the count distributions, as well as the correlations between different subgraph counts.

A first step in this direction is given by considering the expected *squared count*, $\langle n_\gamma^2 \rangle$, for a fixed graph γ . The count itself consists in a sum over embedding positions, and so the squared count is given by summing over two independent embedding positions, which can be reorganized as a sum over the *relative* position of the two copies as defined by vertex and edge coincidences, and a sum over the *absolute* embedding position of the resulting composite graph.

The key point is that the contribution to $\langle n_\gamma^2 \rangle$ from each possible configuration of the composite graph γ_2 is given by *its naive expected count* $\langle n_{\gamma_2} \rangle$ as a subgraph of Γ , multiplied by the number of *distinct ways to combine the two copies into* γ_2 . The multiplication by the number of ways to obtain γ_2 compensates e.g. for the extra twofold symmetry typically arising in γ_2 , related to the interchange of the two copies.

Let us consider the case of a connected γ being a tree or having a single loop, and do a brief analysis of the possible scaling properties of the expected count

$\langle n_{\gamma_2} \rangle$ of the combined graph γ_2 .

For a connected γ , the expected embedding count scales as $O(N^{v-e})$, yielding $O(N)$ for a tree and $O(1)$ for a one-loop graph. When combining two copies of γ into γ_2 , they may overlap in a common subgraph, the *overlap graph*, with edge and vertex counts e_o, v_o . Then the combined graph γ_2 will have vertex count $v_2 = 2v - v_o$ and edge count $e_2 = 2e - e_o$, and its expected count will scale as $O(N^{2v-2e-v_o+e_o})$.

If γ is a *tree*, its only possible overlap graphs are forests, with $v_o - e_o \geq 0$, with equality only for the empty subgraph. This means that the leading $O(N^2)$ contribution to $\langle n_{\gamma}^2 \rangle$ comes entirely from the case where the two copies of γ are completely disjoint, yielding a leading contribution to $\langle n_{\gamma}^2 \rangle$ matching that of $\langle n_{\gamma} \rangle^2$, while the remaining contributions scale at most as $O(N)$. As a result the standard deviation of the γ count scales at most as $O(N^{1/2})$, as compared to the $O(N)$ behavior of the expected count, yielding an asymptotically sharp distribution for the corresponding *intensive* entity, the *count density* $\rho_{\gamma} = n_{\gamma}/N$.

If γ has a *single loop*, we have $v - e = 0$, and the expected count is *finite*. Then we are interested in contributions to $\langle n_{\gamma}^2 \rangle$ scaling at least as $O(1)$, requiring $v_o - e_o \leq 0$. Hence, the only interesting overlap graphs between the two copies of γ are the *empty graph* and connected one-loop graphs (including the entire γ) where the two copies of γ share the loop part (possibly rotated or flipped) both yielding $v_o - e_o = 0$. There are two possibilities here.

If γ consists of a *bare loop* without decorations, the only interesting contributions to $\langle n_{\gamma}^2 \rangle$ are those from cases where the two copies are completely disjoint or completely identical, yielding $\langle n_{\gamma}^2 \rangle = \langle n_{\gamma} \rangle^2 + \langle n_{\gamma} \rangle$ to leading order. The argument can be generalized to higher moments of n_{γ} , showing that the asymptotic distribution of the n_{γ} is *Poissonian* for such γ .

Alternatively, if γ consists of a *decorated loop*, i.e. a single loop with attached tree decorations, there are additional contributions to $\langle n_{\gamma}^2 \rangle$ to leading order, due to configurations of γ_2 where the two copies of γ share the loop but not all of the decorations; as a result the asymptotic distribution of n_{γ} fails to be Poissonian, and is typically wider.

5. Count correlations

In a similar way, the *correlation* between the counts for two distinct small graphs, γ and γ' , say, can be analyzed by considering the expected value of the product of their counts, $\langle n_{\gamma} n_{\gamma'} \rangle$. Again, this can be seen as a sum over their relative embedding positions and over the absolute position of the combined graph.

If both graphs are *trees*, the leading contribution

to $\langle n_{\gamma} n_{\gamma'} \rangle$ comes from cases where the two subgraphs are completely disjoint.

In the *mixed case* of one graph being a tree, the other a connected one-loop graph, the leading contribution to $\langle n_{\gamma} n_{\gamma'} \rangle$ again comes entirely from the completely disjoint case. The argument can be generalized to higher moments, indicating the asymptotic lack of correlations between the two counts.

The final case of interest is when *both* graphs are connected one-loop graphs. If their loops *differ* in length, the leading contribution to $\langle n_{\gamma} n_{\gamma'} \rangle$ again stems entirely from the completely disjoint cases, and the counts are asymptotically uncorrelated. If the loops have the *same* length, however, there are additional contributions from cases where the overlap graph contains the loop, yielding a positive correlation between the two counts.

For a discussion of subgraph counts in the context of the (not necessarily sparse) classic model, based on the concepts of balanced and strictly balanced subgraphs, see e.g. chpt. 4 of ref. [2].

B. Subgraphs statistics II: CDRG-s

Next, we wish to study the statistics of small subgraph counts in a *CDRG-s* ensemble, obtained as the restriction to simple graphs of the corresponding multigraph ensemble, where simple means the absence of loops of length one and two.

Thus, we are led to study the distribution of such loops in the multigraph ensemble, as represented by the subgraph counts when γ is a *pure 1-cycle* (vertex with a tadpole) or a *pure 2-cycle* (two vertices connected by a double edge). The relevant results from the previous subsection as applied to these counts imply the following for a random graph Γ from a CDRG-m ensemble:

- The expected number $\langle n_1 \rangle$ of 1-cycles in Γ is asymptotically given by $\alpha \equiv \text{Tr}(\mathbf{TE})/2$, and the count asymptotically follows a Poissonian distribution, $\text{Prob}(n_1) = e^{-\alpha} \alpha^{n_1}/n_1!$.
- The expected number $\langle n_2 \rangle$ of 2-cycles in Γ is asymptotically given by $\beta \equiv \text{Tr}(\mathbf{TE})^2/4$, and the count asymptotically follows a Poissonian distribution, $\text{Prob}(n_2) = e^{-\beta} \beta^{n_2}/n_2!$.
- The 1-cycle and 2-cycle counts are asymptotically uncorrelated, with each other as well as with the count of any small *simple* graph γ in the form of a tree or a one-loop connected graph.

There are two important implications for the corresponding CDRG-s ensemble:

- The probability that a random graph from the associated multigraph ensemble be simple is asymptotically given by

$$\text{Prob}(\text{simple}) = e^{-\alpha-\beta}, \quad (10)$$

as claimed in ref. [15] without a detailed proof.

- The count distribution for a *simple* small subgraph γ in a random graph drawn from a CDRG-s ensemble is to leading order *asymptotically identical* to the corresponding distribution in a random graph drawn from the corresponding CDRG-m ensemble.

As a result, the computational rules for subgraph counts given in the previous subsection apply without modification also to CDRG-s, for the asymptotically expected subgraph counts of small *simple* graphs to leading order. For a *nonsimple* γ , the count will of course vanish identically – a simple graph has only simple subgraphs.

VI. GLOBAL PROPERTIES

The original CDRG article [15] contained a brief generating function analysis of the asymptotic size distribution of connected components (clusters). Here we give a more detailed derivation, combined with a more elaborate analysis of the result.

A. Connected component statistics

Consider a large random graph Γ drawn from an arbitrary CDRG-m ensemble. Let P_n be the distribution of the number of vertices $0 \leq n < \infty$ of a cluster as revealed by starting from a random vertex in Γ and recursively revealing neighbors of previously revealed vertices. Let $g(z)$ be its generating function,

$$g(z) = \sum_n P_n z^n. \quad (11)$$

At any finite stage in the revelation process, loops in the subgraph revealed so far are suppressed with factors of $1/N$. Thus, in the thermodynamic limit we expect the revealed subgraph, as long as it is finite, to form a tree, and as a result, the following analysis can be expected to apply equally well to the corresponding ensemble of *simple graphs*.

In terms of the generating function $H(\mathbf{x})$ for $\{p_{\mathbf{m}}\}$, as defined in eq. (1), $g(z)$ can be expressed as

$$g(z) = zH(\mathbf{h}(z)) \quad (12)$$

in terms of the set of similarly defined generating functions $h_a(z)$ for the number of vertices in the subtree revealed by following the edge emanating from a random stub of given color a . The rationale behind eq. (12) is that the initial vertex has a random colored degree \mathbf{m} drawn from the distribution $p_{\mathbf{m}}$. This yields a factor z for the initial vertex and a factor $h_a(z)$ for each of its m_a stubs of color a ; summing the result over \mathbf{m} , weighted with $p_{\mathbf{m}}$, yields eq. (12).

The edge functions $h_a(z)$ must satisfy the recursive equations,

$$h_a(z) = z \sum_b T_{ab} \partial_b H(\mathbf{h}(z)), \quad (13)$$

following from a similar argument: An edge emanating from a stub of color a has the color b in the other end with probability $T_{ab} \langle m_b \rangle$, and is then attached to a vertex with colored degree \mathbf{m} with probability $p_{\mathbf{m}} m_b / \langle m_b \rangle$. This yields a total factor of $T_{ab} m_b p_{\mathbf{m}}$. Throw in a factor z to account for that vertex, and a factor $h_c(z)$ for each subtree reached via one of its remaining $m_c - \delta_{cb}$ stubs of color c ; finally, summing over b and \mathbf{m} yields eq. (13).

B. The phase transition and the emergence of the giant

Of particular interest is the result for $z = 1$. The recurrence eq. (13) for $\mathbf{h}(z)$ for the case of $z = 1$ possesses a trivial fixed point $\mathbf{h}(1) = \mathbf{1}$, yielding $g(1) = 1$, expressing the conservation of probability. However, this fixed point represents the physical solution only if it is stable, as determined by the Jacobian \mathbf{J} associated with the linearized recurrence in the neighborhood of the fixed point. \mathbf{J} has the components

$$J_{ab} = \sum_c T_{ac} E_{cb} \Leftrightarrow \mathbf{J} = \mathbf{TE} \quad (14)$$

in terms of the matrix \mathbf{E} of second order combinatorial moments. If all eigenvalues of $\mathbf{J} = \mathbf{TE}$ are less than unity, the trivial fixed point $\mathbf{h}(1) = \mathbf{1}$ is stable, and the revelation asymptotically corresponds to a subcritical branching process, always yielding finite trees.

Otherwise, the trivial fixed point $\mathbf{h}(1) = \mathbf{1}$ is unstable and will repel the iterates of the recursion, eq. (13). This signals that the asymptotic branching process is supercritical, with a finite probability of producing infinite trees. In such a case, a non-trivial fixed point will appear and attract the iterates, yielding a solution with $\mathbf{h}_a(1) < 1$, implying $g(1) < 1$ by virtue of eq. (12). The corresponding probability deficit $1 - g(1)$ is interpreted as being due to the existence of a *giant component*, and measures

the finite probability that the randomly chosen vertex belongs to the giant, asymptotically containing a fraction $1 - g(1)$ of the vertices.

In analogy to the case of a single color, i.e. DRG, the transition is typically second order, being due to an initially unstable, nontrivial fixed point passing the trivial one while they exchange stability characters – a transcritical bifurcation in the language of dynamical systems.

C. Duality

For a *supercritical* model, the solution for $g(z)$ resulting from eq. (12) for the stable fixed point of the recursion, eq. (13), corresponds to a generating function for the contributions from finite clusters only, and can be shown to emulate another, subcritical CDRG model – the *dual* model – as follows. Define properly normalized functions $\hat{g}(z), \hat{h}_a(z)$ in terms of the stable solutions $g(z), \mathbf{h}(z)$ as

$$\hat{h}_a(z) = \frac{h_a(z)}{h_a(1)}, \quad (15a)$$

$$\hat{g}(z) = \frac{g(z)}{g(1)}. \quad (15b)$$

These will then satisfy (by rewriting (12,13))

$$\hat{g}(z) = z\hat{H}(\hat{\mathbf{h}}(z)), \quad (16a)$$

$$\hat{h}_a(z) = z \sum_b \hat{T}_{ab} \partial_b \hat{H}(\hat{\mathbf{h}}(z)), \quad (16b)$$

where $\hat{H}(\mathbf{x})$ and $\hat{\mathbf{T}}$ are given by

$$\hat{H}(\mathbf{x}) = \frac{H(\mathbf{h}(1) \times \mathbf{x})}{H(\mathbf{h}(1))}, \quad (17a)$$

$$\hat{T}_{ab} = \frac{1}{h_a(1)} T_{ab} \frac{1}{h_b(1)}. \quad (17b)$$

They describe the *dual* CDRG model, that is subcritical by definition: The stable fixed point is mapped to $\hat{\mathbf{h}}(1) = \mathbf{1} \Rightarrow \hat{g}(1) = 1$. The corresponding transformed CDD is obtained from the original one by a geometric transform, $\hat{p}_{\mathbf{m}} \propto p_{\mathbf{m}} \mathbf{h}(1)^{\mathbf{m}}$. This duality has analogues in other sparse models, such as DRG (trivially), IRG [8], and the classic model [2].

VII. REDUNDANCY

A CDRG model defines a unique ensemble of graphs. The opposite is not generally true – there is a built-in redundancy in the CDRG description, such that several models may describe one and the same graph ensemble, as we will now demonstrate, based on the local as well as the global properties

of the graphs in a CDRG ensemble, as analyzed in sections V and VI above.

Consider a given asymptotic CDRG model, and define a transformed model by using a stochastic matrix \mathbf{U} , $\mathbf{U}\mathbf{1} = \mathbf{1}$, to define transformed H and \mathbf{T} as

$$\hat{H}(\mathbf{x}) = H(\mathbf{U}\mathbf{x}), \quad (18a)$$

$$\hat{\mathbf{T}} = \mathbf{U}^{-1} \mathbf{T} \mathbf{U}^{-\top}. \quad (18b)$$

This transform conserves the EDD normalization, $H(\mathbf{1}) = 1$, and leaves form-invariant the constraint on \mathbf{T} , eq. (3).

It also leaves invariant the recursive relations, eq. (13), for the generating functions $\mathbf{h}(z)$ for the size of a subtree found by following an edge starting from a stub of definite color, if $\mathbf{h}(z)$ is transformed to $\hat{\mathbf{h}}(z) = \mathbf{U}^{-1} \mathbf{h}(z)$. This leaves $g(z)$ invariant by virtue of eq. (12), and thus will not affect the observable distribution of component sizes, $\{P_n\}$.

As for the local properties in the form of expected small subgraph counts, also these are left invariant, since the computational (Feynman) rules given in Section V A 2 invariably yield expressions in the form of *contractions* between the color indices of combinatorial moments $\mathbf{E}_{(n)}$ on the one hand, and those of the color preference matrix \mathbf{T} on the other.

The transform can be interpreted as a *change of basis* in color space, such that U_{ab} gives the probability $P_{b|a}$ that the original color a corresponds to the transformed color b .

This suggests the existence of a continuous symmetry group, $\sim \text{SL}(K-1)$, for the class of CDRG models. Of course, we have to be careful to stay in the physical regime, with non-negative values for $\{T_{ab}\}$ and $\{p_{\mathbf{m}}\}$ (but not necessarily for $\{U_{ab}\}$ themselves), which restricts the possible transforms and prevents the class of transformations to form a group. Nevertheless, it implies that CDRG consists in *equivalence classes* of models, related by transformations of the type (18a,18b).

One can consider even more general transformations, where also the number of colors, K , is changed, requiring a non-square \mathbf{U} . This enables the reducibility under certain conditions of a model to an equivalent model with a smaller color space.

VIII. SUBCLASSES EQUIVALENT TO OTHER MODELS

A. DRG

The restriction of CDRG to a single color, $K = 1$, trivially yields DRG, where a plain degree distribution $\{p_m\}$ is given, while \mathbf{T} reduces to a number T , constrained to equal $\langle m \rangle^{-1}$ by virtue of eq. (3).

More generally, a DRG model effectively results as soon as \mathbf{T} has rank one, in which case \mathbf{T} takes the form of a direct product, forced to equal $\mathbf{T}_{\text{DRG}} = \mathbf{1} \langle m \rangle^{-1} \mathbf{1}^\top$, with all components equal. This prevents the stub colors from affecting the stub pairing statistics, resulting in a completely random, unbiased stub pairing.

B. IRG

Next we wish to identify the CDRG subclass corresponding to IRG. To that end, consider the restriction of CDRG to ensembles of simple graphs with a colored degree distribution given by a mixture of multivariate Poissonians,

$$p_{\mathbf{m}} = \sum_{i=1}^L r_i \prod_a \exp(-C_{ia}) C_{ia}^{m_a} / m_a! \quad (19)$$

equivalent to

$$H(\mathbf{x}) = \sum_{i=1}^L r_i \exp(\mathbf{C}_i \cdot (\mathbf{x} - \mathbf{1})) \quad (20)$$

for some $L \geq 1$, where each term in the sum over i corresponds to a non-negative weight r_i times a normalized multivariate Poissonian with colored degree average $\langle \mathbf{m} \rangle_i = \mathbf{C}_i = \{C_{ia}\}$, with the weights summing up to unity, $\sum_i r_i = 1$.

In ref. [15] the asymptotic equivalence of such an ensemble to an associated IRG ensemble was shown, based on an analysis of the equations (12,13) for the cluster size distribution. IRG (inhomogeneous random graphs) [8] is defined as a colored extension of the classic model of simple graphs, where a distinct ensemble of graphs of size N is defined in terms of colored vertices, where each vertex is independently assigned a color $i \in [1 \dots L]$ according to an arbitrary but fixed distribution $\{r_i\}$. Then for every pair of vertices, the corresponding edge is independently realized with a color-dependent probability given by c_{ij}/N , where i, j are the colors assigned to the vertices.

For such a model, a generating function analysis of the cluster size distribution can be done, analogous to the one represented by eqs. (12,13). The result [8] is that for IRG, the generating function for the cluster size distribution, $g(z)$ as defined in eq. (11), can be written as a weighted sum

$$g(z) = \sum_i r_i g_i(z) \quad (21)$$

where $g_i(z)$ is the generating function for the size distribution, *conditional* on the IRG vertex color i

of a randomly chosen initial vertex. These satisfy a set of recursive relations amounting to

$$g_i(z) = z \exp \left(\sum_j c_{ij} r_j (g_j(z) - 1) \right) \quad (22)$$

As shown in ref. [15], by defining $g_i(z) = z \exp(\sum_a C_{ia}(h_a(z) - 1))$ and $c_{ij} = \sum_{ab} C_{ia} T_{ab} C_{jb}$, eqs (12,13) can be written in the form of eqs. (21, 22), showing the asymptotic equivalence from the point of view of cluster size distributions.

An interesting question then is whether this relation persists when considering small subgraph counts. In the rules for the computation of the asymptotically expected count of a small subgraph γ , as defined in section V, each vertex in γ with n stubs is associated with a factor $N \mathbf{E}_{(n)}$. For a CDD as defined by eq. (20), the combinatorial moment $\mathbf{E}_{(n)}$ simplifies to $\sum_{i=1}^L r_i \mathbf{C}_i^{\circ n}$, where $\mathbf{C}_i^{\circ n}$ stands for the outer (tensor) product of n factors of \mathbf{C}_i , one for every stub. Absorbing these stub factors into the edge factor, \mathbf{T}/N , yields a set of Feynman rules with an independent IRG color i for each vertex, acquiring a corresponding factor $N r_i$, and a factor of $\mathbf{C}_i^\top \mathbf{T} \mathbf{C}_j / N = c_{ij}/N$ for every edge connecting a pair of vertices with respective IRG colors i, j . The product of vertex and edge factors should be summed over the IRG colors i, j, \dots , and the result divided by the usual symmetry factor S_γ .

Indeed, these are the correct rules for the expected simple subgraph counts in an IRG model, as can be derived using simple arguments; this confirms the asymptotic equivalence between the two models previously indicated by the cluster size analysis.

C. Other subclasses

As a special case of the IRG subclass, a CDRG-s ensemble with a CDD in the form of a single multivariate Poissonian, as defined by eq. (20) with a single term, $H(\mathbf{x}) = \exp(\mathbf{C} \cdot (\mathbf{x} - \mathbf{1}))$, is asymptotically equivalent to the classic model with the parameter value $c = \mathbf{C}^\top \mathbf{T} \mathbf{C} = \mathbf{C}^\top \cdot \mathbf{1} = \sum_a C_a$.

Another interesting subclass of CDRG is defined by the restriction to models with *monochrome* vertices [10], such that for each vertex all its stubs are forced to have the same color. It is a trivial exercise to derive the rules for subgraph counts as well as the equations for the generating function $g(z)$ for the cluster size distribution in such an ensemble.

In fact, the monochrome subclass is sufficient for spanning IRG, since for a given IRG model as defined by $\{c_{ij}\}, \{r_i\}$, one can always find an associated CDRG model with the identical color space by using a *diagonal* matrix, $C_{ia} = C_i \delta_{ia}$ with $C_i =$

$\sum_j c_{ij} r_j$. This yields an equivalent monochrome CDRG model defined by $T_{ij} = c_{ij}/(C_i C_j)$ and $H(\mathbf{x}) = \sum_i r_i \exp(C_i(x_i - 1))$.

IX. CONCLUDING REMARKS

We have considered and analyzed a recently suggested general class of ensembles, CDRG, of sparse random graphs, based on a hidden coloring of stubs. We have extended the formalism to incorporate ensembles of multigraphs (CDRG-m), in addition to the originally considered ensembles of simple graphs (CDRG-s).

A distinct random graph model can be defined asymptotically by specifying a colored degree distribution $\{p_{\mathbf{m}}\}$, controlling the distribution in the number and colors of the connections of a node, and a color preference matrix \mathbf{T} , governing the relative tendency for connections between stubs with definite pairs of colors. Based on such an asymptotic model, an ensemble of simple graphs or multigraphs of a given size can be defined.

For such models, we have demonstrated the calculability of local as well as global observable structural properties, important for the anticipated use of the formalism as a target for model inference based on the observed properties of real-world networks.

Local graph characteristics can be represented by the statistics of small subgraph counts. We have derived a set of simple rules for calculating the asymptotically expected count of an arbitrary small graph, and demonstrated the equivalence between the two types of ensembles (of simple or multigraphs) as far as simple subgraphs counts are concerned. We have also discussed the shapes of the count distributions, and shown that a Poissonian distribution results asymptotically only for simple cycles. By comparing the expected counts in DRG and CDRG of certain simple subgraphs, we have demonstrated the role of the hidden coloring in enabling a non-trivial edge correlation structure.

Global properties have been exemplified by the statistics of cluster sizes, for which we have performed a detailed analysis using generating function techniques. The analysis shows that an arbitrary CDRG model displays a percolation threshold at a well-defined critical hypersurface in parameter space, above which a giant component appears containing a finite fraction of the vertices in the thermodynamic limit. We have also demonstrated for a

supercritical model the existence of a dual model – an associated subcritical model describing the non-giant part.

The algebraic properties of the equations involved in both the local and global analysis reveal a redundancy (or symmetry) in CDRG, such that several superficially distinct models describe the same observable ensemble of graphs. This redundancy can be seen as being due to the possibility of a change of basis in the abstract color space.

The rules for the computation of expected subgraph counts have a form strongly reminiscent of Feynman rules for perturbative calculations in statistical field theory, indicating a relationship between CDRG models and field theories, in analogy to the case for DRG [20]. Work is in progress to explore such relations, and the results will be presented in a separate article [21].

The CDRG class of random graph models is very general, and contains several previously studied models and classes of models as special cases. Its structure is also such that it should admit a straightforward generalization e.g. to models of directed graphs. While CDRG so far has been considered only for degree distributions with exponential fall-off for large degrees, it should be extendable to power-behaved degree distributions if proper care is taken. The key obstacle (inherited from DRG) is that in such a case the higher moments of the (colored) degree distribution diverge, which makes some observables – in particular for CDRG-s – very sensitive to the precise definition of the ensembles.

Anticipating that the formalism can be extended as indicated above, a few fundamental questions remain to be answered. **(1)** Is the resulting class “*complete*”, i.e. does it span every reasonable model of sparse, truly random graphs? If not, how generalize it? **(2)** Is it unnecessarily general, i.e. can an arbitrary CDRG model be reformulated in a simple way entirely in terms of observable graph properties, without utilizing hidden variables such as color?

Acknowledgments

An informative discussion with K. Nowicki on the complications associated with subgraph distributions is gratefully acknowledged. This work was in part supported by the Swedish Foundation for Strategic Research.

[1] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17 (1960).

[2] B. Bollobás, *Random Graphs, 2nd ed.* (Cambridge University Press, Cambridge, 2001).

- [3] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs* (Wiley and Sons, New York, 2000).
- [4] P. Flajolet, D. E. Knuth, and B. Pittel, *Discr. Math.* **75**, 167 (1989).
- [5] E. A. Bender and E. A. Canfield, *J. Combinat. Theory A* **24**, 296 (1978).
- [6] M. Molloy and B. Reed, *Rand. Struct. Alg.* **6**, 161 (1995).
- [7] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- [8] B. Söderberg, *Phys. Rev. E* **66**, 066121 (2002).
- [9] J. Berg and M. Lässig, *Phys. Rev. Lett.* **89**, 228701 (2002).
- [10] M. E. J. Newman, *Phys. Rev. E* **67**, 026126 (2003).
- [11] R. Albert and A.-L. Barabási, *Phys. Rev. Lett.* **85**, 5234 (2000).
- [12] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz, *Phys. Rev. E* **64**, 041902 (2001).
- [13] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, *Phys. Rev. E* **63**, 062101 (2001).
- [14] T. S. Turova, *Phys. Rev. E* **65**, 066102 (2002).
- [15] B. Söderberg (2003), arXiv:cond-mat/0303466.
- [16] T. Łuczak, in *Poznań, 1989*, edited by A. Frieze and T. Łuczak (John Wiley & Sons, New York, 1992), vol. 2 of *Random Graphs*, pp. 165–182.
- [17] M. Molloy and B. Reed, *Combinat. Prob. Comput.* **7**, 295 (1998).
- [18] D. J. Watts and S. H. Strogatz, *Nature* **393/4**, 440 (1998).
- [19] S. Maslov and K. Sneppen, *Science* **296**, 910 (2002).
- [20] Z. Burda, J. D. Correia, and A. Krzywicki, *Phys. Rev. E* **64**, 046118 (2001).
- [21] B. Söderberg (work in progress).
- [22] Due to the restriction to even M , the generating function is really given by $(H(\mathbf{z})^N + H(-\mathbf{z})^N) / (H(\mathbf{1})^N + H(-\mathbf{1})^N)$, yielding an unimportant modification. We will use $H(\mathbf{z})^N$ for simplicity, keeping in mind that only the even- M part is physical.