The Grand Canonical Ensemble of Weighted Networks

Andrea Gabrielli,1, 2 Rossana Mastrandrea,2 Guido Caldarelli,2, 1 and Giulio Cimini2, 1

1 Istituto dei Sistemi Complessi (CNR) UoS Sapienza, P.le A. Moro 2, 00185 Rome (Italy)
2 IMT School for Advanced Studies, Piazza San Francesco 19, 55100 Lucca (Italy)

The cornerstone of statistical mechanics of complex networks is the idea that the links, and not the nodes, are the effective particles of the system. Here we formulate a mapping between weighted networks and lattice gasses, making the conceptual step forward of interpreting weighted links as particles with a generalized coordinate. This leads to the definition of the grand canonical ensemble of weighted complex networks. We derive exact expressions for the partition function and thermodynamic quantities, both in the cases of global and local (i.e., node-specific) constraints on density and mean energy of particles. We further show that, when modeling real cases of networks, the binary and weighted statistics of the ensemble can be disentangled, leading to a simplified framework for a range of practical applications.

What distinguishes a network from the systems typically studied in physics is the complex heterogeneous pattern of interactions (links) among its constituent elements (nodes). Indeed, the statistical mechanics approach to networks has been developed treating the interactions themselves as the degrees of freedom of the system, pushing forward the interpretation of links as the actual particles of the system [1]. Under this view, the maximum number of particles \( V \) (i.e., the maximum number of links) is the equivalent of the volume of a physical system. For binary networks with fixed number \( N \) of nodes, a number of seminal works [2, 3] defined the canonical ensemble by fixing the number of links \( L \), and the grand canonical ensemble by letting \( L \) fluctuate around its expected value. For instance in the Erdős-Rényi model, these two cases correspond to the Erdős-Rényi model, these two cases correspond to the binary adjacency matrix \( A \approx \sum_{i,j} \delta(i,j) \) and the set of weights \( \{ w_{ij} \} \) associated to them (meaning that only existing links \( (i,j) \in L \subseteq V \) with \( |L| = L \) (i.e., the set of node pairs with \( a_{ij} = 1 \) and the set of weights \( \{ w_{ij} \} \) associated to them (meaning that only existing links/particles contribute to the statistics of the system)). Therefore, the binary adjacency matrix \( A \approx \sum_{i,j} \delta(i,j) \) and the set of weights \( \{ w_{ij} \} \) associated to them (meaning that only existing links/particles contribute to the statistics of the system). Therefore, the binary adjacency matrix \( A \approx \sum_{i,j} \delta(i,j) \) and the set of weights \( \{ w_{ij} \} \) associated to them (meaning that only existing links/particles contribute to the statistics of the system).

The information entropy associated to the probability distribution is \( P(C) = P(A, W) \equiv P(L, \{ w_{ij} \}) \), and the sum over configuration is performed as

\[
\sum_{C} = \sum_{A} \prod_{i<j} \int_{0}^{\infty} dw_{ij}.
\]

The information entropy associated to the probability measure \( P(C) \) is as usual \( S = -\sum_{C} P(C) \log P(C) \), and the shape of \( P(C) \) is found by maximizing \( S \) under given constraints. This is the framework of Exponential Ran-
The equations determining the values of $\alpha$ and $\beta$ are then

$$
\langle L \rangle \equiv -\partial_{\alpha} \log Z_G(\alpha, \beta) \equiv \frac{V}{\beta e^\alpha + 1} = L^*, 
$$

$$
\langle W \rangle \equiv -\partial_{\beta} \log Z_G(\alpha, \beta) \equiv \frac{V \beta^{-1}}{\beta e^\alpha + 1} = W^*,
$$

from which we immediately find $\beta^{-1} = W^*/L^* = w^*$, i.e., the mean weight, and $1 + e^\alpha/w^* = V/L^*$. We thus see that while $\beta$ controls for the mean weight (energy) of existing links (particles), $\alpha$ controls for the mean density of links (particles). Note that since between each pair of nodes there can be only a single link/particle, the system can be represented with $V$ copies of a Fermi system having a single energy level $\varepsilon = 1$. Under this analogy, $\log \beta$ plays the role of the inverse absolute temperature $(kT)^{-1}$, whereas, $-\alpha$ is the ratio $\mu (kT)^{-1}$ between chemical potential and temperature.

Remarkably, we can perform the parameter transformation $\alpha' = \alpha + \log \beta$, so that $\alpha'$ alone determines the mean link density and, given this density, $\beta$ alone sets the mean weight of existing links: we have

$$
P(A, W) = \left[ \prod_{i<j} \frac{e^{-\alpha' a_{ij}}}{1 + e^{-\alpha}} \right] \left[ \frac{\prod_{i<j} \beta e^{-\beta w_{ij}}}{\beta} \right].
$$

This shows that individual link occupations are all mutually independent events and that, given a binary configuration $A$, weight values of individual existing links are also independent events. Besides, moments of link occupation and of link weight probability distributions can be independently set in order to satisfy the constraints. As explicitly shown in the next section, this property is due to the global nature of the constraints. Note that for equilibrium statistical mechanics with short range interactions, if the system is homogeneous then local and global measures coincide.

Local constraints — We now impose for each node $i$ the mean degree or number of incident links $\langle k_i \rangle \equiv (\sum_{j \neq i} a_{ij}) = k_i^*$ and the mean strength or total weight of incident links $\langle s_i \rangle = (\sum_{j \neq i} w_{ij}) = s_i^*$. This grand canonical ensemble is analogous to the continuous version of the enhanced configuration model [12], whence we use the acronym CECM. We have $P(A, W, \{\alpha_i, \beta_i\}_{i=1}^N) = Z_G^{-1} \left\{ \{\alpha_i, \beta_i\}_{i=1}^N \right\} e^{-H(A, W, \{\alpha_i, \beta_i\}_{i=1}^N)}$ with

$$
H(A, W, \{\alpha_i, \beta_i\}_{i=1}^N) = \sum_{i<j} (\alpha_i + \alpha_j) a_{ij} + \sum_{i<j} (\beta_i + \beta_j) w_{ij},
$$

$$
Z_G \left\{ \{\beta_i\}_{i=1}^N \right\} = \sum_A e^{-H(A, W, \{\alpha_i, \beta_i\}_{i=1}^N)} = \sum_A e^{-\sum_{i<j} (\alpha_i + \alpha_j) a_{ij}} Z_C \left( \beta_i \right)_{i=1}^N
$$

and $Z_C \left( \beta_i \right)_{i=1}^N = \prod_{i<j} (\beta_i + \beta_j)^{-1}$. Performing the sum over all binary configurations leads to

![Diagram](image.png)
\[ Z_G (\{\alpha_i, \beta_i\}_{i=1}^N) = \sum_A \frac{e^{-\sum_{i<j} (\alpha_i + \alpha_j) a_{ij}}}{\prod_{i<j} (\beta_i + \beta_j)} = \sum_A \prod_{i<j} \frac{e^{-(\alpha_i + \alpha_j) a_{ij}}}{\beta_i + \beta_j} = 1 + \sum_{U \subset V} \prod_{i<j} \frac{e^{-(\alpha_i + \alpha_j)}}{\beta_i + \beta_j} = \prod_{i<j} \left( 1 + \frac{e^{-(\alpha_i + \alpha_j)}}{\beta_i + \beta_j} \right) \]  

(10)

where \( U \) is a generic subset of \( V \). The values of the multipliers are then found through the constraints equations:

\[ \langle k_i \rangle \equiv -\partial_{\alpha_i} \log Z_G (\{\alpha_i, \beta_i\}_{i=1}^N) \equiv \sum_{j \neq i} \frac{1}{1 + e^{-(\beta_i + \beta_j) e^{\alpha_i + \alpha_j}} = k_i^s, \]  

(11)

\[ \langle s_i \rangle \equiv -\partial_{\beta_i} \log Z_G (\{\alpha_i, \beta_i\}_{i=1}^N) \equiv \sum_{j \neq i} \frac{(\beta_i + \beta_j)^{-1}}{1 + (\beta_i + \beta_j) e^{\alpha_i + \alpha_j}} = s_i^s, \]  

(12)

\( \forall i \). Note that after some algebra we can rewrite \( P(A, W) \) as

\[ P(A, W) = \left[ \prod_{i<j} \frac{e^{-(\alpha_i + \alpha_j + \log(\beta_i + \beta_j) | a_{ij}}}{1 + e^{-(\alpha_i + \alpha_j + \log(\beta_i + \beta_j))} \right] \left[ \prod_{i<j} (\beta_i + \beta_j) e^{-(\beta_i + \beta_j) w_{ij}} \right] = \pi(A) q(W) \]  

(13)

with \( \pi(A) \) being the unconditional probability distribution of the binary configuration \( A \), and \( q(W) \) the probability density function of the weights of the existing links (i.e., the set \( \mathcal{L} \)) conditional to \( A \). The form of \( q(W) \) is exponential, differently from the geometric and Poissonian forms obtained in [12] and [14] respectively, due to the continuous nature of the weights.

At this point some considerations are in order. I) Both \( \pi(A) \) and \( q(W) \) factorize into the product of single link probability distributions: occupations of different links are independent events and, conditional to the binary configuration, weights of different links are also independent. II) However the parameters defining single link probabilities and weights are entangled, which means that local link densities cannot be set independently on local weights, because of the simultaneous conservations of mean node degrees and strengths. Such an interplay allows better clarifying the role of nodes (and in particular of node heterogeneity), which play the role of interactions between links (particles). Indeed only if node properties are homogeneous, like when we impose global constraints, such topological interactions disappear: the system is spatially homogeneous in terms of density of particles and of energy, which can be thus set independently. The statistical mechanical case analogous to a heterogeneous network situation instead arises when we constrain the local mean particle and energy densities, \( n(x) \) and \( \varepsilon(x) \), to be homogeneous, i.e., both dependent on \( x \). This case is typically not encountered in ordinary equilibrium statistical mechanics, with the possible exception of glassy disordered systems and long range interactions. III) If we look at the generic link occupation probability

\[ p_{ij} = \frac{e^{-(\alpha_i + \alpha_j + \log(\beta_i + \beta_j))}}{1 + e^{-(\alpha_i + \alpha_j + \log(\beta_i + \beta_j))}} \]  

(14)

from the viewpoint of statistical mechanics, we can again interpret the single link problem as a single state local Fermi system with energy level \( \varepsilon = 1 \), inverse local temperature \( (kT_{ij})^{-1} = \log(\beta_i + \beta_j) \), and local chemical potential \( \mu_{ij} = -kT_{ij}(\alpha_i + \alpha_j) \). However, differently from the homogeneous case, different links are not independent copies of the same problem, but topologically interacting single-level Fermi systems with different local temperatures and chemical potentials—which are mutually related by local heterogeneous constraints.

**Separability of binary and weighted statistics** — We finally explore the separability of local links densities and weights distribution also for the case of local constraints [21]. To this end we introduce a two-step entropy maximization procedure, the *separable* enhanced configuration model (SECM):

1. We first constrain the mean node degrees only, obtaining the probability of the binary configuration \( A \) as for the standard configuration model [3]:

\[ \pi(A) = \prod_{i<j} \frac{e^{-(\alpha_i + \alpha_j)| a_{ij}}}{1 + e^{-(\alpha_i + \alpha_j))} \]  

(15)

2. Then, for each \( A \), we constrain the mean node strengths, obtaining the probability density of the link weights conditional to \( A \) (coinciding with that of the CECM):

\[ q(W_A) = \prod_{i<j} (\beta_i + \beta_j) e^{-(\beta_i + \beta_j) w_{ij}}. \]  

(16)

The SECM is thus defined by the constraint equations

\[ \langle k_i \rangle \equiv \sum_{j \neq i} \frac{1}{1 + e^{\alpha_i + \alpha_j}} = k_i^s, \]  

(17)

\[ \langle s_i \rangle \equiv \sum_{j \neq i} \frac{(\beta_i + \beta_j)^{-1}}{1 + e^{\alpha_i + \alpha_j}} = s_i^s, \]  

(18)
FIG. 2. Properties of CECM and SECM ensembles in four real networks: the World Trade Web, the eMID interbank network, the neural network of C. elegans, and the human functional brain network. The upper part of the figure shows the comparison of link probabilities (first row) and of expected weights (second row) obtained by CECM and SECM. The lower part of the figure instead shows how the two ensembles reproduce higher-order statistics of the real networks (defined in): nearest neighbors degree $k_{nn}$ (third row), clustering coefficient $c$ (fourth row), nearest neighbors strength $s_{nn}$ (fifth row) and weighted clustering $c_{w}$ (sixth row).
∀i, and by the joint probability distribution

\[ P(A, W) = \prod_{i<j} e^{-(\alpha_i' + \alpha_j') s_{ij}} \prod_{i<j} \left[ \frac{1}{1 + e^{-(\alpha_i' + \alpha_j')}} \right] \prod_{i<j} (\beta_i + \beta_j) e^{-(\beta_i + \beta_j) w_{ij}} \]  \tag{19}  

By definition, in the SECM the parameters defining link probabilities and weights are disentangled, so that the local statistics for these quantities can be set independently. In the CECM instead the parameters controlling for link weights also play a role in determining connection strengths, differently. In the CECM instead the parameters controlling for link weights also play a role in determining connection strengths and degrees satisfy

\[
\gamma_k = \log(\gamma_k + \beta_0) \forall k.
\]

Fig. 2). However, in the CECM connection probabilities, the contribution of parameters \( \{\beta_i\}_{i=1}^N \) is logarithmic with respect to that of parameters \( \{\alpha_i\}_{i=1}^N \): weighted properties in general give only small perturbations to the Lagrange multipliers of node degrees. As such, CECM and SECM define similar link probabilities and expected weights (Fig. 2), and are almost interchangeable for all practical purposes—the advantage of SECM being an easier numerical implementation. Finally, it is noteworthy that CECM and SECM coincide when the constraints on strengths and degrees satisfy \( s_i^* = \gamma k_i^* \forall i \) for constant \( \gamma \). Indeed in this case \( \beta_i = \beta_0 \forall i \), and thus we have the exact correspondence \( \alpha_i' = \alpha_i + \frac{1}{2} \log(2\beta_0) \forall i \). This is for instance the case of HFBN of Fig. 2.

**Final remarks** — Ensembles of random graphs with given structural properties like those discussed here typically find a twofold application [11][22]. On one hand, they can be taken as null model networks and thus be used to assess the significance of patterns observed for a real network. On the other hand, when details on the microscopic structure of a real network are unknown, they can be used to reconstruct the most likely network configuration. The grand canonical ensemble introduced here represents, both in its rigorous version and separable approximation, a very versatile tool for these tasks, being defined for the most general class of networks with continuous weights. For instance, the *fitness-induced configuration model* [20] used to reconstruct networks without degree information is easily implemented in our grand canonical framework [22]. Extensions of the framework to the case of multiplex networks [24] as well as to include higher order interactions between generalized coordinates will be covered in future work.

**Acknowledgements** — A.G. acknowledges support from the CNR PNR project CRISIS-Lab. R.M. G.C. and G.C. acknowledge support from the EU H2020 projects DOLFINS (grant 640772), SoBigData (grant 654024) and OpenMaker (grant 687941).