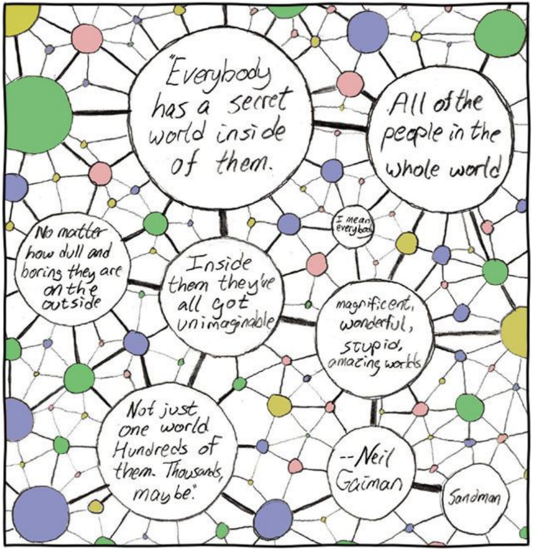
Network Science

2019. november 18.



THE HIDDEN VARIABLE MODEL

The hidden variable model

- How to generate a random graph with degree correlations?

→ With the hidden variable model!

The hidden variable model

Definition of the hidden variable model

- We define a distribution $\rho(h)$ for the hidden variables, and a joint distribution $r(h, h')$ for the linking probabilities.

The hidden variable model

Definition of the hidden variable model

- We define a distribution $\rho(h)$ for the hidden variables, and a joint distribution $r(h, h')$ for the linking probabilities.
- We take N nodes, and for each we draw its hidden variable from $\rho(h)$.

The hidden variable model

Definition of the hidden variable model

- We define a distribution $\rho(h)$ for the hidden variables, and a joint distribution $r(h, h')$ for the linking probabilities.
- We take N nodes, and for each we draw its hidden variable from $\rho(h)$.
- Every node pair i and j are connected with a probability

$$P(i - j) = r(h_i, h_j).$$

The hidden variable model

- Let $P(k \mid h)$ denote the conditional probability that a node has degree k given its hidden variable is h.

The hidden variable model

- Let $P(k \mid h)$ denote the conditional probability that a node has degree k given its hidden variable is h.
- Based on that the degree distribution can be written as

$$p(k) = \sum_h P(k \mid h)\rho(h)$$

The hidden variable model

- Let $P(k \mid h)$ denote the conditional probability that a node has degree k given its hidden variable is h.
- Based on that the degree distribution can be written as

$$p(k) = \sum_h P(k \mid h)\rho(h)$$

- Normalisation:

$$\sum_k P(k \mid h) = 1$$

  for all h.

The hidden variable model

- Let $P(k \mid h)$ denote the conditional probability that a node has degree k given its hidden variable is h.
- Based on that the degree distribution can be written as

$$p(k) = \sum_h P(k \mid h)\rho(h)$$

- Normalisation:

$$\sum_k P(k \mid h) = 1$$

  for all h.
- The average degree of nodes with a given h can be written as

$$\overline{k(h)} = \sum_k kP(k \mid h)$$

The hidden variable model

- Let $P(k \mid h)$ denote the conditional probability that a node has degree k given its hidden variable is h.
- Based on that the degree distribution can be written as

$$p(k) = \sum_h P(k \mid h)\rho(h)$$

- Normalisation:

$$\sum_k P(k \mid h) = 1$$

  for all h.
- The average degree of nodes with a given h can be written as

$$\overline{k(h)} = \sum_k kP(k \mid h)$$

  and the overall average degree as

$$\langle k \rangle = \sum_k kp(k) = \sum_k k \sum_h P(k \mid h)\rho(h) = \sum_h \overline{k(h)}\rho(h).$$

The hidden variable model

- Let us unfold $P(k \mid h)$ as

$$P(k \mid h) = \sum_{k_1,\ldots,k_c} P_1^{(h)}(k_1, h_1) P_2^{(h)}(k_2, h_2) \cdots P_c^{(h)}(k_c, h_c) \delta_{k_1+k_2+\cdots+k_c, k},$$

  where $P_i^{(h)}(k_i, h_i)$ denotes the probability that a node with hidden variable h is connected altogether with k_i links to other nodes with hidden variable h_i, and h_c is the maximal value of h.

The hidden variable model

- Let us unfold $P(k \mid h)$ as

$$P(k \mid h) = \sum_{k_1,\ldots,k_c} P_1^{(h)}(k_1, h_1) P_2^{(h)}(k_2, h_2) \cdots P_c^{(h)}(k_c, h_c) \delta_{k_1+k_2+\cdots+k_c,k},$$

  where $P_i^{(h)}(k_i, h_i)$ denotes the probability that a node with hidden variable h is connected altogether with k_i links to other nodes with hidden variable h_i, and h_c is the maximal value of h.
- Since links are introduced independent of each other,

$$P_i^{(h)}(k_i, h_i) = \binom{N_i}{k_i} r(h_i, h)^{k_i} \left[1 - r(h_i, h)\right]^{N_i - k_i},$$

  where $N_i = N\rho(h_i)$ is the number of nodes with hidden variable h_i.

The hidden variable model

- Let us unfold $P(k \mid h)$ as

$$P(k \mid h) = \sum_{k_1,\ldots,k_c} P_1^{(h)}(k_1,h_1)P_2^{(h)}(k_2,h_2)\cdots P_c^{(h)}(k_c,h_c)\delta_{k_1+k_2+\cdots+k_c,k},$$

  where $P_i^{(h)}(k_i,h_i)$ denotes the probability that a node with hidden variable h is connected altogether with k_i links to other nodes with hidden variable h_i, and h_c is the maximal value of h.

- Since links are introduced independent of each other,

$$P_i^{(h)}(k_i,h_i) = \binom{N_i}{k_i} r(h_i,h)^{k_i}\left[1-r(h_i,h)\right]^{N_i-k_i},$$

  where $N_i = N\rho(h_i)$ is the number of nodes with hidden variable h_i.

- The **generating function** of this binomial distribution is

$$G_i^{(h)}(z) = \sum_{k_i} \binom{N_i}{k_i} r(h_i,h)^{k_i}\left[1-r(h_i,h)\right]^{N_i-k_i} z^{k_i} = \left[1-(1-z)r(h_i,h)\right]^{N_i}.$$

The hidden variable model

- Since $k = k_1 + k_2 + \cdots + k_c$ and $P(k \mid h)$ is given by the convolution of the $P_i^{(h)}(k_i, h_i)$ distributions, the generating function of $P(k \mid h)$, denoted by $G(z)$ is simply

$$G(z) = \prod_i G_i^{(h)}(z) = \prod_i \left[1 - (1 - z) r(h_i, h)\right]^{N_i}.$$

The hidden variable model

- Since $k = k_1 + k_2 + \cdots + k_c$ and $P(k \mid h)$ is given by the convolution of the $P_i^{(h)}(k_i, h_i)$ distributions, the generating function of $P(k \mid h)$, denoted by $G(z)$ is simply

$$G(z) = \prod_i G_i^{(h)}(z) = \prod_i \left[1 - (1 - z) r(h_i, h)\right]^{N_i} .$$

- If we take the logarithm of the generating function, we get

$$\begin{aligned} \ln G(z) &= \sum_i N_i \ln\left[1 - (1 - z) r(h_i, h)\right] = \\ & N \sum_{h'} \rho(h') \ln\left[1 - (1 - z) r(h', h)\right] . \end{aligned}$$

The hidden variable model

- Since $k = k_1 + k_2 + \cdots + k_c$ and $P(k \mid h)$ is given by the convolution of the $P_i^{(h)}(k_i, h_i)$ distributions, the generating function of $P(k \mid h)$, denoted by $G(z)$ is simply

$$G(z) = \prod_i G_i^{(h)}(z) = \prod_i \left[1 - (1 - z) r(h_i, h)\right]^{N_i} .$$

- If we take the logarithm of the generating function, we get

$$\begin{aligned} \ln G(z) \quad &= \quad \sum_i N_i \ln\left[1 - (1 - z) r(h_i, h)\right] = \\ & \quad\; N \sum_{h'} \rho(h') \ln\left[1 - (1 - z) r(h', h)\right] . \end{aligned}$$

- The average degree of a node with hidden variable h is given by $G'(z)|_{z=1}$.

The hidden variable model

- Since $k = k_1 + k_2 + \cdots + k_c$ and $P(k \mid h)$ is given by the convolution of the $P_i^{(h)}(k_i, h_i)$ distributions, the generating function of $P(k \mid h)$, denoted by $G(z)$ is simply

$$G(z) = \prod_i G_i^{(h)}(z) = \prod_i \left[1 - (1 - z) r(h_i, h)\right]^{N_i}.$$

- If we take the logarithm of the generating function, we get

$$\begin{aligned} \ln G(z) &= \sum_i N_i \ln\left[1 - (1 - z) r(h_i, h)\right] = \\ & N \sum_{h'} \rho(h') \ln\left[1 - (1 - z) r(h', h)\right]. \end{aligned}$$

- The average degree of a node with hidden variable h is given by $G'(z)\big|_{z=1}$. However, since $G(1) = 1$, this is also equal to

$$\begin{aligned} \overline{k(h)} &= G'(z)\big|_{z=1} = \left[\ln G(z)\right]'\big|_{z=1} = N \sum_{h'} \left.\frac{\rho(h') r(h', h)}{1 - (1 - z) r(h', h)}\right|_{z=1} = \\ & N \sum_{h'} \rho(h') r(h', h). \end{aligned}$$

The hidden variable model

- Let us assume that we want to generate **sparse** networks, where $\overline{k(h)}$ **is independent of N**.

The hidden variable model

- Let us assume that we want to generate **sparse** networks, where **$k(h)$ is independent of N**.

→ If $\rho(h)$ is size independent as well (which is a quite natural assumption), then $r(h',h)$ must behave as

$$r(h',h) = \frac{C(h',h)}{N},$$

where $C(h',h)$ is a size independent "nice" function.

The hidden variable model

- Let us assume that we want to generate **sparse** networks, where $\overline{k(h)}$ **is independent of** N.

→ If $\rho(h)$ is size independent as well (which is a quite natural assumption), then $r(h', h)$ must behave as

$$r(h', h) = \frac{C(h', h)}{N},$$

where $C(h', h)$ is a size independent "nice" function.

- Based on that, the logarithm of the generating function can be written as

$$\begin{aligned}
\ln G(z) &= N \sum_{h'} \rho(h') \ln\left[1 - (1-z) r(h', h)\right] = \\
& N \sum_{h'} \rho(h') \ln\left[1 - (1-z)\frac{C(h', h)}{N}\right] \simeq \\
& N \sum_{h'} \rho(h') \left[(z-1)\frac{C(h', h)}{N}\right],
\end{aligned}$$

The hidden variable model

- Let us assume that we want to generate **sparse** networks, where $\overline{k(h)}$ **is independent of** N.

→ If $\rho(h)$ is size independent as well (which is a quite natural assumption), then $r(h',h)$ must behave as

$$r(h',h) = \frac{C(h',h)}{N},$$

where $C(h',h)$ is a size independent "nice" function.

- Based on that, the logarithm of the generating function can be written as

$$\begin{aligned}
\ln G(z) \quad &= \quad N\sum_{h'}\rho(h')\ln\left[1-(1-z)r(h',h)\right] = \\
&\quad\; N\sum_{h'}\rho(h')\ln\left[1-(1-z)\frac{C(h',h)}{N}\right] \simeq \\
&\quad\; N\sum_{h'}\rho(h')\left[(z-1)\frac{C(h',h)}{N}\right],
\end{aligned}$$

yielding for the generating function itself

$$G(z) \simeq \exp\left[(z-1)\sum_{h'}\rho(h')C(h',h)\right] = \exp\left[(z-1)\overline{k(h)}\right].$$

The hidden variable model

- The probability mass function corresponding to an exponential generating function is the Poisson distribution, thus,

$$
\begin{aligned}
P(k \mid h) &= \frac{\overline{k(h)}^{k} e^{-\overline{k(h)}}}{k!}, \\
\overline{k(h)} &= N \sum_{h'} \rho(h') r(h', h) = \sum_{h'} \rho(h') C(h', h).
\end{aligned}
$$

The hidden variable model

- What about the **degree correlations**?

The hidden variable model

- What about the **degree correlations**?

→ Let's start with defining $P(h' \mid h)$ as the conditional probability that a link is connecting to a node with hidden variable h', given that the node at the other end has h.

The hidden variable model

- What about the **degree correlations**?

→ Let's start with defining $P(h' \mid h)$ as the conditional probability that a link is connecting to a node with hidden variable h', given that the node at the other end has h.

- Since the number of such links is proportional to both the number of nodes with h' and the connection probability $r(h', h)$, we can write

$$P(h' \mid h) = \frac{\rho(h')r(h',h)}{\sum_{h''} \rho(h'')r(h'',h)} = \frac{N\rho(h')r(h',h)}{\overline{k(h)}}.$$

The hidden variable model

- What about the **degree correlations**?

→ Let's start with defining $P(h' \mid h)$ as the conditional probability that a link is connecting to a node with hidden variable h', given that the node at the other end has h.

- Since the number of such links is proportional to both the number of nodes with h' and the connection probability $r(h', h)$, we can write

$$P(h' \mid h) = \frac{\rho(h')r(h',h)}{\sum_{h''}\rho(h'')r(h'',h)} = \frac{N\rho(h')r(h',h)}{\overline{k(h)}}.$$

- Based on $P(h' \mid h)$ we can write the conditional probability $P(k' \mid k)$ that a link connects to a node with degree k, given that it started on a node with degree k as

$$P(k' \mid k) = \sum_{h',h} P(k' \mid h')P(h' \mid h)P(h \mid k),$$

where the last factor $P(h \mid k)$ can be given according to Bayes theorem as

$$P(h \mid k) = \frac{P(k \mid h)\rho(h)}{\sum_k P(k \mid h)\rho(h)} = \frac{P(k \mid h)\rho(h)}{p(k)}.$$

The hidden variable model

- By substituting back we obtain

$$P(k' \mid k) = \frac{1}{p(k)} \sum_{h',h} P(k' \mid h') P(h' \mid h) P(k \mid h) \rho(h).$$

The hidden variable model

- By substituting back we obtain

$$P(k' \mid k) = \frac{1}{p(k)} \sum_{h',h} P(k' \mid h') P(h' \mid h) P(k \mid h) \rho(h).$$

- Thus, the ANND can be writen as

$$\begin{aligned} k_{\mathrm{nn}}(k) &= \sum_{k'} k' P(k' \mid k) = \frac{1}{p(k)} \sum_{k',h',h} \underbrace{k' P(k' \mid h')}_{\overline{k'(h')}} P(h' \mid h) P(k \mid h) \rho(h) \\ & \frac{1}{p(k)} \sum_{h',h} \underbrace{\overline{k'(h')} P(h' \mid h)}_{k_{\mathrm{nn}}(h)} P(k \mid h) \rho(h), \end{aligned}$$

where we introduced

$$k_{\mathrm{nn}}(h) = \sum_{h'} \overline{k'(h')} P(h' \mid h)$$

as the average nearest neighbours degree of nodes with parameter h.

The hidden variable model

- By substituting back we obtain

$$P(k' \mid k) = \frac{1}{p(k)} \sum_{h',h} P(k' \mid h')P(h' \mid h)P(k \mid h)\rho(h).$$

- Thus, the ANND can be writen as

$$\begin{aligned} k_{\mathrm{nn}}(k) &= \sum_{k'} k'P(k' \mid k) = \frac{1}{p(k)} \sum_{k',h',h} \underbrace{k'P(k' \mid h')}_{\overline{k'(h')}} P(h' \mid h)P(k \mid h)\rho(h) \\ & \frac{1}{p(k)} \sum_{h',h} \underbrace{\overline{k'(h')}P(h' \mid h)}_{k_{\mathrm{nn}}(h)} P(k \mid h)\rho(h), \end{aligned}$$

where we introduced

$$k_{\mathrm{nn}}(h) = \sum_{h'} \overline{k'(h')}P(h' \mid h)$$

as the average nearest neighbours degree of nodes with parameter h. Thus, the ANND can be formulated as

$$k_{\mathrm{nn}}(k) = \frac{1}{p(k)} \sum_{h} k_{\mathrm{nn}}(h)P(k \mid h)\rho(h).$$

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

- Let us go back to the complete statistical description of degree correlations:

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

- Let us go back to the complete statistical description of degree correlations:
  - The matrix $E_{k',k}$ gives the number of connections between nodes of degree k and k',

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

- Let us go back to the complete statistical description of degree correlations:
  - The matrix $E_{k',k}$ gives the number of connections between nodes of degree k and k',
  - The matrix $e_{k',k} = E_{k',k}/2M$ gives the probability for finding a link between nodes of degree k and k', and can be treated as a **joint probability mass function** between k and k'.

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

- Let us go back to the complete statistical description of degree correlations:
  - The matrix $E_{k',k}$ gives the number of connections between nodes of degree k and k',
  - The matrix $e_{k',k} = E_{k',k}/2M$ gives the probability for finding a link between nodes of degree k and k', and can be treated as a **joint probability mass function** between k and k'.
  - The degree distribution can be calculated from $e_{k',k}$ as

$$q_k = \sum_{k'} e_{k',k} = \frac{kp(k)}{\langle k \rangle} \quad \longrightarrow \quad p(k) = \frac{\langle k \rangle}{k} q_k = \frac{\langle k \rangle}{k} \sum_{k'} e_{k',k}.$$

The hidden variable model

- OK, but how to generate a random graph with prescribed degree correlations?

→ We can use the method of **hidden degrees**.

- Let us go back to the complete statistical description of degree correlations:
  - The matrix $E_{k',k}$ gives the number of connections between nodes of degree k and k',
  - The matrix $e_{k',k} = E_{k',k}/2M$ gives the probability for finding a link between nodes of degree k and k', and can be treated as a **joint probability mass function** between k and k'.
  - The degree distribution can be calculated from $e_{k',k}$ as

$$q_k = \sum_{k'} e_{k',k} = \frac{kp(k)}{\langle k \rangle} \quad \longrightarrow \quad p(k) = \frac{\langle k \rangle}{k} q_k = \frac{\langle k \rangle}{k} \sum_{k'} e_{k',k}.$$

- Analogously, if we specify a **joint probability mass function** $P(h', h)$, the distribution of the hidden variables can be written as

$$\rho(h) = \frac{\langle h \rangle}{h} \sum_{h'} P(h', h).$$

The hidden variable model

- Since the number of nodes with h is $N_h = N\rho(h)$, it is natural to set the connection probabilities as

$$r(h', h) = \frac{\langle h \rangle P(h', h)}{N\rho(h)\rho(h')}.$$

The hidden variable model

- Since the number of nodes with h is $N_h = N\rho(h)$, it is natural to set the connection probabilities as

$$r(h',h) = \frac{\langle h \rangle P(h',h)}{N\rho(h)\rho(h')}.$$

- This way, the average degree of nodes with h is

$$\overline{k(h)} = N\sum_{h'} r(h',h)\rho(h') = \langle h \rangle \sum_{h'} \frac{P(h',h)}{\rho(h)} = h,$$

The hidden variable model

- Since the number of nodes with h is $N_h = N\rho(h)$, it is natural to set the connection probabilities as

$$r(h', h) = \frac{\langle h \rangle P(h', h)}{N\rho(h)\rho(h')}.$$

- This way, the average degree of nodes with h is

$$\overline{k(h)} = N \sum_{h'} r(h', h)\rho(h') = \langle h \rangle \sum_{h'} \frac{P(h', h)}{\rho(h)} = h,$$

- and therefore

$$P(k \mid h) = \frac{h^k e^{-h}}{k!}.$$

The hidden variable model

- Since the number of nodes with h is $N_h = N\rho(h)$, it is natural to set the connection probabilities as

$$r(h', h) = \frac{\langle h \rangle P(h', h)}{N\rho(h)\rho(h')}.$$

- This way, the average degree of nodes with h is

$$\overline{k(h)} = N \sum_{h'} r(h', h)\rho(h') = \langle h \rangle \sum_{h'} \frac{P(h', h)}{\rho(h)} = h,$$

- and therefore

$$P(k \mid h) = \frac{h^k e^{-h}}{k!}.$$

- Thus, the degree distribution and the ANND are

$$\begin{aligned} p(k) &= \sum_h \frac{h^k e^{-h}}{k!} \rho(h), \\ k_{nn}(k) &= \frac{1}{p(k)} \sum_{h'} k_{nn}(h) \frac{h^k e^{-h}}{k!} \rho(h), \end{aligned}$$

where

$$k_{nn}(h) = \sum_{h'} \overline{k'(h')} P(h' \mid h) = \sum_{h'} h' P(h' \mid h).$$

The hidden variable model

- The $P(h' \mid h)$ conditional probability can be obtained from the joint distribution $P(h', h)$ as

$$P(h' \mid h) = \frac{e_{h',h}}{\sum_{h'} e_{h',h}} = \frac{e_{h',h}}{q_h} = \frac{\langle h \rangle \, P(h',h)}{h\rho(h)}.$$

The hidden variable model

- The $P(h' \mid h)$ conditional probability can be obtained from the joint distribution $P(h', h)$ as

$$P(h' \mid h) = \frac{e_{h',h}}{\sum_{h'} e_{h',h}} = \frac{e_{h',h}}{q_h} = \frac{\langle h \rangle P(h', h)}{h\rho(h)}.$$

- Since the Poisson distribution is a narrow distribution around its mean, in the large degree regime

$$\begin{aligned} p(k) &\sim \rho(h = k), \\ k_{\mathrm{nn}}(k) &\sim k_{\mathrm{nn}}(h = k) \end{aligned}$$