EURANDOM PREPRINT SERIES

2015-011

July 1, 2015

Complexe netwerken vanuit fysisch perspectief

D. Garaschelli, F. den Hollander, A. Roccaverde
ISSN 1389-2355

D. Garlaschelli
Instituut-Lorentz voor Theoretische Fysica
Universiteit Leiden
Niels Bohrweg 2
2333 CA Leiden
garlaschelli@lorentz.leidenuniv.nl

F. den Hollander
Mathematisch Instituut
Universteit Leiden
Postbus 9512
2300 RA Leiden
denholla@math.leidenuniv.nl

A. Roccaverde
Mathematisch Instituut
Universteit Leiden
Postbus 9512
2300 RA Leiden
a.roccaverde@math.leidenuniv.nl

Complexe netwerken vanuit fysisch perspectief

Voor veeldeeltjessystemen geldt in de regel dat de *mikro-kanonieke* beschrijving in termen van energie equivalent is aan de *kanonieke* beschrijving in termen van temperatuur, in de thermodynamische limiet wanneer het aantal deeltjes naar oneindig gaat. Echter, wanneer de deeltjes over lange afstanden met elkaar wisselwerken, dan kan deze equivalentie worden gebroken. Voor complexe netwerken – grote toevallige grafen met topologische randcondities – doet zich eenzelfde verschijnsel voor. In dit artikel geven we diverse voorbeelden en kwantificeren we de mate van niet-equivalentie met behulp van het begrip entropie.

1. Statistische fysica

In de statistische fysica wordt voor het bestuderen van de eigenschappen van veeldeeltjessystemen in evenwicht gebruik gemaakt van zogenaamde *ensembles*, d.w.z. kansverdelingen op de verzameling van deeltjesconfiguraties (Boltzmann [1]). Welk ensemble wordt gekozen hangt af van de informatie die over het systeem beschikbaar is (Gibbs [3]). Zo correspondeert het *mikro-kanonieke ensemble* met de uniforme kansverdeling op de deelverzameling van al die deeltjesconfiguraties waarvoor de energie een voorgeschreven waarde heeft. De keuze voor de uniforme kansverdeling drukt daarbij uit dat van het systeem geen andere informatie beschikbaar is dan dat het de voorgeschreven energie heeft, zodat elke configuratie met deze energie een a priori gelijke kans heeft. Echter, het rekenen met een 'harde randconditie' is in het algemeen lastig. In de praktijk is het handiger om te werken met een 'zachte randconditie'. Het *kanonieke ensemble* correspondeert met de kansverdeling met maximale entropie op de verzameling van alle deeltjesconfiguraties, zonder restrictie op de energie, maar zodanig dat de *gemiddelde* energie een voorgeschreven waarde heeft. De keuze voor de maximale entropie drukt daarbij uit dat van het systeem geen andere informatie beschikbaar is dan dat het de voorgeschreven gemiddelde energie heeft. Er wordt een geschikte *temperatuur* gekozen, die wiskundig gezien de rol speelt van een Lagrange multiplicator die de zachte randconditie realiseert. Het kanonieke ensemble kan gezien worden als de oplossing van een probleem van *statistische inferentie* op basis van de *partiële informatie* die verpakt zit in deze randconditie (Jaynes [6]).

Voor systemen waarin de deeltjes over *korte afstanden* met elkaar wisselwerken is aangetoond dat de beide ensembles equivalent zijn in de thermodynamische limiet, d.w.z. wanneer het aantal deeltjes naar oneindig gaat. Het idee hier achter is dat in het kanonieke ensemble de energie weliswaar fluctueert, maar op een schaal die verwaarloosbaar klein is ten opzichte van de gemiddelde energie, zodat het zich effectief gedraagt als het mikro-kanonieke ensemble. Er zijn echter voorbeelden van systemen waarin de deeltjes over *lange afstanden* met elkaar wisselwerken waarvoor de equivalentie wordt gebroken. De achtergrond van dit verschijnsel is nog niet goed begrepen.

Vanuit wiskundig perspectief zijn er diverse manieren om niet-equivalentie te kwantificeren (Ellis, Touchette en Turkington [2], [10]). Een recente suggestie is om te kijken naar de *relatieve entropie van de twee ensembles per deeltje* en te laten zien dat die niet naar nul convergeert in de thermodynamische limiet (Touchette [11]). Relatieve entropie is een pseudo-afstand op de ruimte van kansverdelingen en kan derhalve kansverdelingen van elkaar onderscheiden. Onder bepaalde aannamen kan worden aangetoond dat deze vorm van kwantificeren van niet-equivalentie de scherpste is.

2. **Intermezzo: relatieve entropie**
Zij χ een eindige verzameling. Laten $\vec{p} = (p_i)_{i\in\chi}$ en $\vec{q} = (q_i)_{i\in\chi}$ kansverdelingen zijn op χ. De relatieve entropie van $\vec{p}$ t.o.v. $\vec{q}$ is gedefinieerd als

$$S(\vec{p} \mid \vec{q}) = \sum_{i\in\chi} p_i \log\left(\frac{p_i}{q_i}\right).$$

Omdat de functie $x \mapsto h(x) = x \ln x$ strict convex is op $[0,\infty)$, geldt dat

$$\begin{aligned} S(\vec{p} \mid \vec{q}) &= \sum_{i\in\chi} q_i\, h\left(\frac{p_i}{q_i}\right) \geq h\left(\sum_{i\in\chi} q_i \left(\frac{p_i}{q_i}\right)\right) \\ &= h\left(\sum_{i\in\chi} p_i\right) = h(1) = 0. \end{aligned}$$

M.a.w. de relatieve entropie is niet-negatief. Merk op dat gelijkheid geldt dan en slechts dan wanneer $\vec{p} = \vec{q}$. Merk verder op dat $S(\vec{p} \mid \vec{q})$ niet symmetrisch is onder verwisseling van $\vec{p}$ en $\vec{q}$.

3. **Complexe netwerken.**
In dit artikel bestuderen we ensemble equivalentie voor *complexe netwerken*: grote toevallige grafen met topologische randcondities (Squartini, de Mol, den Hollander en Garlaschelli [9]). De rol van de deeltjesconfiguratie wordt daarbij overgenomen door de graaf zelf, en de energie van de deeltjesconfiguratie door een geschikte functie van de graaf.

Voor $N \in \mathbb{N}$, zij $\mathcal{G}_N$ de verzameling van alle grafen met N knooppunten. Zij $\vec{C}$ een vector-waardige functie op $\mathcal{G}_N$. De *mikro-kanonieke kansverdeling* met *harde randconditie* $\vec{C}^\star$ is gedefinieerd als

$$P_{\text{mic}}(\mathbf{G}) = \begin{cases} 1/\Omega_{\vec{C}^\star}, & \vec{C}(\mathbf{G}) = \vec{C}^\star, \\ 0, & \vec{C}(\mathbf{G}) \neq \vec{C}^\star, \end{cases} \qquad \mathbf{G} \in \mathcal{G}_N,$$

waar

$$\Omega_{\vec{C}^\star} = |\{\mathbf{G} \in \mathcal{G}_N \colon\ \vec{C}(\mathbf{G}) = \vec{C}^\star\}|$$

het aantal grafen is dat $\vec{C}^\star$ realiseert. De *kanonieke kansverdeling* wordt gedefinieerd als de kansverdeling op $\mathcal{G}_N$ die de *Shannon entropy*

$$S_N(P_{\text{can}}) = -\sum_{\mathbf{G}\in\mathcal{G}_N} P_{\text{can}}(\mathbf{G}) \ln P_{\text{can}}(\mathbf{G})$$

maximaliseert onder de *zachte randconditie* $\langle \vec{C} \rangle = \vec{C}^\star$, waarbij $\langle\cdot\rangle$ het gemiddelde is m.b.t. P_{can}. Deze 'duale' kansverdeling wordt gegeven door

$$P_{\text{can}}(\mathbf{G}) = \frac{\exp[-H(\mathbf{G},\vec{\theta}^\star)]}{Z(\vec{\theta}^\star)}, \qquad \mathbf{G} \in \mathcal{G}_N,$$

waar

$$H(\mathbf{G},\vec{\theta}) = \vec{\theta} \cdot \vec{C}(\mathbf{G})$$

de zogenaamde *Hamilton-functie* op $\mathcal{G}_N$ is en

$$Z(\vec{\theta}) = \sum_{\mathbf{G}\in\mathcal{G}_N} \exp[-H(\mathbf{G},\vec{\theta})]$$

de normalisatieconstante. De parameter $\vec{\theta}$ moet gelijk worden gekozen aan de waarde $\vec{\theta}^\star$ waarvoor $\langle \vec{C} \rangle = \vec{C}^\star$. Deze waarde maximaliseert de 'statistical likelihood' van de 'random data' $\mathbf{G}$.

4. **Specifieke relatieve entropie van ensembles**
De relatieve entropie van P_{mic} t.o.v. P_{can} is

$$S_N(P_{\text{mic}} \mid P_{\text{can}}) = \sum_{\mathbf{G}\in\mathcal{G}_N} P_{\text{mic}}(\mathbf{G}) \ln \frac{P_{\text{mic}}(\mathbf{G})}{P_{\text{can}}(\mathbf{G})}.$$

Net als in Touchette [11], zeggen we dat de twee ensembles equivalent zijn dan en slechts dan als de *specifieke relatieve entropie* nul is, d.w.z.

$$s = \lim_{N\to\infty} \frac{S_N(P_{\text{mic}} \mid P_{\text{can}})}{N} = 0.$$

Vanwege de vorm van $H(\mathbf{G},\vec{\theta})$ geldt dat $P_{\text{can}}(\mathbf{G}_1) = P_{\text{can}}(\mathbf{G}_2)$ zodra $\vec{C}(\mathbf{G}_1) = \vec{C}(\mathbf{G}_2)$. De kanonieke kans is dus hetzelfde voor alle configuraties met dezelfde waarde van de randconditie. Derhalve geldt dat

$$S_N(P_{\text{mic}} \mid P_{\text{can}}) = \ln \frac{P_{\text{mic}}(\mathbf{G}^\star)}{P_{\text{can}}(\mathbf{G}^\star)},$$

waarbij $\mathbf{G}^\star$ een *willekeurige* graaf in $\mathcal{G}_N$ is waarvoor $\vec{C}(\mathbf{G}^\star) = \vec{C}^\star$. Ensemble equivalentie betekent dus dat

$$\lim_{N\to\infty} \frac{1}{N}\big[\ln P_{\text{mic}}(\mathbf{G}^\star) - \ln P_{\text{can}}(\mathbf{G}^\star)\big] = 0,$$

m.a.w. het asymptotisch gedrag van de kans van die configuraties die aan de harde randconditie voldoen is voor beide ensembles hetzelfde (den Hollander [5]). Deze observatie is niet alleen van theoretisch belang, het vereenvoudigt ook aanzienlijk de berekeningen, zoals we zullen laten zien.

5. **Configuratie Model**
Het *Configuratie Model* genereert een toevallige graaf bestaande uit N punten met van te voren vastgelegde graden $\vec{k}^\star = (k_1^\star,\dots,k_N^\star)$, waarbij $k_i^\star \in \mathbb{N}_0 = \{0,1,2,\dots\}$ het aantal lijnen is dat met knooppunt i is verbonden. In dit voorbeeld is de randconditie dus $\vec{C}^\star = \vec{k}^\star$. Het aantal grafen $\Omega_{\vec{k}^\star}$ dat aan deze randconditie voldoet is niet bekend, maar er zijn wel asymptotische resultaten voor het geval dat $N \to \infty$ en

$$(\star) \qquad k_{\max} = \max_{1\leq i\leq N} k_i^\star = o(\sqrt{N}).$$

Namelijk, onder conditie $(\star)$ geldt dat (McKay and Wormald [7])

$$\Omega_{\vec{k}^\star} = \frac{\sqrt{2}\,(\frac{2L^\star}{e})^{L^\star}}{\prod_{i=1}^N k_i^\star!}\, e^{-[\overline{k^{\star 2}}/2\overline{k^\star}]^2 + \frac{1}{4} + o(N^{-1}\overline{k^\star}^{\,3})},$$

waar

$$\begin{array}{lcll} \overline{k^\star} & = & N^{-1}\sum_{i=1}^{N} k_i^\star & \text{(gemiddelde graad)}, \\ \overline{k^{\star 2}} & = & N^{-1}\sum_{i=1}^{N} k_i^{\star 2} & \text{(gemiddelde kwadratische graad)}, \\ L^\star & = & \frac{1}{2}N\overline{k^\star} & \text{(totaal aantal lijnen)}. \end{array}$$

De kanonieke kansverdeling correspondeert met $H(\mathbf{G},\vec{\theta}) = \vec{\theta}\cdot\vec{k}(\mathbf{G})$, en $\vec{\theta}^\star$ is de oplossing van de vergelijking

$$\sum_{j\neq i}\frac{e^{-\theta_i^\star-\theta_j^\star}}{1+e^{-\theta_i^\star-\theta_j^\star}} = k_i^\star \quad \forall\, i$$

(Squartini en Garlaschelli [8]). Met de afkorting $p_{ij}^\star = e^{-\theta_i^\star-\theta_j^\star}/(1+e^{-\theta_i^\star-\theta_j^\star})$ hebben we dan

$$P_{\mathrm{can}}(\mathbf{G}) = \prod_{i=1}^{N}\prod_{j<i}(p_{ij}^\star)^{g_{ij}}(1-p_{ij}^\star)^{1-g_{ij}},$$

waar g_{ij} het (i,j)-element is van de zogenaamde nabuurmatrix van de graaf $\mathbf{G}$ (d.w.z. $g_{ij}=1$ wanneer i en j verbonden zijn door een lijn, en $g_{ij}=0$ anders). De conditie in $(\star)$ garandeert dat $k_{\max}=o(\sqrt{L})$, zodat

$$p_{ij}^\star \sim e^{-\theta_i^\star-\theta_j^\star} = \frac{k_i^\star k_j^\star}{2L^\star} = o(1),$$

waar $\sim$ betekent dat het quotient van de linker- en de rechterzijde naar 1 convergeert. Derhalve geldt dat $\theta_i^\star \sim -\ln(k_i^\star/\sqrt{2L^\star})$, en een eenvoudige berekening leert dat

$$\ln P_{\mathrm{can}}(\mathbf{G}^\star) \sim \sum_{i=1}^{N} k_i^\star \ln k_i^\star - L^\star\ln(2L^\star) - L^\star.$$

Dit leidt op zijn beurt tot de formule

$$S_N(P_{\mathrm{mic}}\mid P_{\mathrm{can}}) \sim \sum_{i=1}^{N} g(k_i^\star) + \left[\,\overline{k^{\star 2}}/2\overline{k^\star}\,\right]^2 - \tfrac{1}{4} + o\left(N^{-1}\overline{k^\star}^{\,3}\right)$$

met

$$g(k) = \ln\left(\frac{k!}{k^k e^{-k}}\right).$$

De *empirische kansverdeling van de graden* is

$$f_N^\star = N^{-1}\sum_{i=1}^{N}\delta_{k_i^\star},$$

waarbij δ_k de punt-kansverdeling is gedefinieerd door

$$\delta_k(l) = 1_{\{l=k\}}, \qquad l\in\mathbb{N}_0.$$

Onder de veronderstelling dat f_N voor $N\to\infty$ convergeert naar een kansverdeling f op $\mathbb{N}_0$, en wel zodanig dat $\lim_{N\to\infty}\sum_{k\in\mathbb{N}_0}|f_N^\star(k)-f(k)|g(k) = 0$ (convergentie in $\ell_1(g)$), vinden we dat

$$s = \sum_{k\in\mathbb{N}_0} f(k)g(k).$$

Omdat $g(k)>0$ voor alle $k\in\mathbb{N}_0\backslash\{0\}$, volgt dat $s>0$ zodra $f\neq\delta_0$, en dus zijn de ensembles niet-equivalent. M.a.w. in de limiet $N\to\infty$ geldt dat onder de kanonieke kansverdeling de meeste grafen *niet* een gradenrij hebben die in de buurt ligt van de gemiddelde gradenrij.

De laatste formule heeft een interessante interpretatie. Namelijk,

$$g(k) = s\big(\delta_k \mid \mathrm{POI[k]}\big)$$

is de relatieve entropie van de punt-kansverdeling δ_k t.o.v. de Poisson-kansverdeling POI[k] gedefinieerd door

$$\mathrm{POI}[k](l) = e^{-k}\,\frac{k^l}{l!}, \qquad l\in\mathbb{N}_0.$$

Dit zegt dat elk knooppunt met graad k een relatieve entropie $g(k)$ bijdraagt aan de totale relatieve entropie. M.a.w. de totale relatieve entropie is een *lineaire functie* van het aantal knooppunten dat aan een (harde dan wel zachte) randconditie op de graad moet voldoen. Het is daarom dat de specifieke relatieve entropie in het Configuratie Model strikt positief is.

6. Twee voorbeelden

We bekijken twee bijzondere gevallen.

(a) **(Reguliere netwerken)** Elk punt heeft dezelfde graad, d.w.z. $k_i^\star = k^\star$ met $k^\star = o(\sqrt{N})$. Dan geldt

$$s = g(k^\star).$$

Merk op dat wanneer $k^\star = k^\star(N)$ divergeert als $N\to\infty$, er een extreme vorm van niet-equivalentie optreedt omdat $g(k^\star)\sim\ln(\sqrt{2\pi k^\star})$, $k^\star\to\infty$.

(b) **(Schaalvrije netwerken)** We kiezen de gradenrij $\vec{k}^\star$ zodanig dat

$$f_N^\star(k) = \begin{cases} A_{k_c}k^{-\gamma}, & 1\le k<k_c, \\ 0, & k\ge k_c, \end{cases}$$

met $\gamma\in(1,\infty)$ en met $k_c = k_c(N)$ zodanig gekozen dat $\lim_{N\to\infty}k_c(N)=\infty$ en $k_c(N)=o(\sqrt{N})$. Wanneer we $f_N^\star$ benaderen door een continue kansverdeling, dan vinden we dat $A_{k_c}\approx\gamma-1$. Omdat $g(k)\approx\ln\sqrt{2\pi k}$ leidt dit tot de approximatieve formule

$$s \approx \frac{1}{2(\gamma-1)} + \ln\sqrt{2\pi}.$$

Merk op dat wanneer γ afneemt de mate van niet-equivalentie toeneemt. Hoe vaker knooppunten met grote graad voorkomen ('hubs' genoemd), hoe sterker de niet-equivalentie is.

7. **Bi-partite grafen**

Gegeven twee verzamelingen van knooppunten U en V, ter grootte M en N. Voor de eerste verzameling leggen we de gradenrij vast op $\vec{k}^\star = (k_1^\star, \dots, k_M^\star)$, voor de tweede op $\vec{l}^\star = (l_1^\star, \dots, l_N^\star)$. Lijnen zijn alleen toegestaan tussen knooppunten in U enerzijds en V anderzijds. In het bijzonder geldt dat

$$\sum_{i=1}^{M} k_i^\star = \sum_{j=1}^{N} l_j^\star = L^\star,$$

met $L^\star$ het totaal aantal lijnen. Zij $\mathcal{G}_{M,N}$ de verzameling van alle bi-partite grafen. Voor P_{mic} kiezen we weer de uniforme verdeling op de deelverzameling van alle grafen in $\mathcal{G}_{M,N}$ die gradenrijen $\vec{k}^\star$ en $\vec{l}^\star$ hebben. Voor P_{can} kiezen we de kansverdeling

$$P_{\mathrm{can}}(\mathbf{G}) = \frac{\exp[-H(\mathbf{G}, \vec{\theta}^\star, \vec{\zeta}^\star)]}{Z(\vec{\theta}^\star, \vec{\zeta}^\star)}, \qquad \mathbf{G} \in \mathcal{G}_{M,N},$$

met Hamilton-functie

$$H(\mathbf{G}, \vec{\theta}^\star, \vec{\zeta}^\star) = \vec{\theta}^\star \cdot \vec{k}(\mathbf{G}) + \vec{\zeta}^\star \cdot \vec{l}(\mathbf{G})$$

en normalisatieconstante $Z(\vec{\theta}^\star, \vec{\zeta}^\star)$, waarbij $\vec{\theta}^\star$ en $\vec{\zeta}^\star$ zo dienen te worden gekozen dat de gemiddelde gradenrijen onder P_{can} gelijk zijn aan $\vec{k}^\star$ en $\vec{l}^\star$.

Onder de veronderstelling dat $L^\star \to \infty$ en

$$(\star) \qquad k_{\max} l_{\max} = o\big(L^{\star\,2/3}\big), \qquad k_{\max} = \max_{1 \le i \le M} k_i^\star, \quad l_{\max} = \max_{1 \le j \le N} l_j^\star,$$

geldt (Greenhill, McKay and Wang [4])

$$\Omega_{\vec{k}^\star, \vec{l}^\star} = \frac{L^\star!}{\prod_{i=1}^{M} k_i^\star! \prod_{j=1}^{N} l_j^\star!} \exp[o(M+N)], \quad M, N \to \infty.$$

(Een meer precieze uitdrukking is beschikbaar voor de exponent, maar die is niet nodig voor onze berekening.) Met behulp van de formula

$$S_{M,N}(P_{\mathrm{mic}} \mid P_{\mathrm{can}}) = \ln \frac{P_{\mathrm{mic}}(\mathbf{G}^\star)}{P_{\mathrm{can}}(\mathbf{G}^\star)},$$

waarbij $\mathbf{G}^\star$ een *willekeurige* graaf in $\mathcal{G}_{M,N}$ is met gradenrijen $\vec{k}^\star$ en $\vec{l}^\star$, leidt dit uiteindelijk tot de uitdrukking

$$\begin{aligned} s &= \lim_{M,N \to \infty} \frac{S_{M,N}(P_{\mathrm{mic}} \mid P_{\mathrm{can}})}{M+N} \\ &= \alpha \sum_{k \in \mathbb{N}_0} f_1(k) g(k) + (1-\alpha) \sum_{k \in \mathbb{N}_0} f_2(k) g(k), \end{aligned}$$

waarbij f_1, f_2 de limieten zijn van de empirische kansverdelingen $f_{1,M}^\star, f_{2,N}^\star$ van de graden, de convergentie naar deze limieten geschiedt in de $\ell_1(g)$-norm, en de limiet $M, N \to \infty$ zo genomen wordt dat

$$\lim_{M,N \to \infty} \frac{M}{M+N} = \alpha \in [0,1].$$

De berekening is analoog aan die voor het Configuratie Model. Opnieuw zien we dat de ensembles niet-equivalent zijn, en dat de totale relatieve entropie op te vatten is als een som over knooppunten van de relatieve entropie per knooppunt.

8. **Conclusie**

Niet-equivalentie van ensembles is een natuurlijke eigenschap van complexe netwerken. De achterliggende reden is dat *locale randcondities een lange-dracht effect* hebben wanneer hun aantal extensief is in het aantal knooppunten. Het is een uitdaging om een volledige *classificatie* te geven van alle randcondities die leiden tot niet-equivalentie.

Referenties

[1] L. Boltzmann, Wiener Berichte 2, 373, 1877.
[2] R. S. Ellis, H. Touchette, and B. Turkington, Physica A 335, 518, 2004.
[3] J. W. Gibbs, *Elementary Principles of Statistical Mechanics*, Yale University Press, New Haven, Connecticut, 1902.
[4] C. Greenhill, B. D. McKay, X. Wang, J. Combin. Theory A 113, 291, 2004.
[5] F. den Hollander, *Large Deviations*, Fields Institute Monographs 14, American Mathematical Society, Providence RI, 2000.
[6] E. T. Jaynes, Phys. Rev. 106, 4, 1957.
[7] B. D. McKay, N. C. Wormald, Combinatorica 11, 369, 1991.
[8] T. Squartini, D. Garlaschelli, New Journal of Physics, 13, 083001, 2011.
[9] T. Squartini, J. de Mol, F. den Hollander, D. Garlaschelli, Breaking of ensemble equivalence in networks, [arXiv:1501.00388].
[10] H. Touchette, R. S. Ellis, B. Turkington, Physica A 340, 138, 2004.
[11] H. Touchette, [arXiv:1403.6608].