UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

LPTMS Laboratoire de Physique Théorique et Modèles Statistiques

Université Paris-Sud

École Doctorale 564 : Physique en Île-de-France
Laboratoire de Physique Théorique et Modèles Statistiques

Discipline : Physique

Thèse de doctorat
Synthèse

Soutenue le 14 Novembre 2014 par

Andrey Y. Lokhov

Méthode de cavité dynamique et problèmes sur des graphes

Devant le jury composé de :

Satya Majumdar	Président du jury
Marc Mézard	Directeur de thèse
David Saad	Rapporteur
Guilhem Semerjian	Examinateur
Lenka Zdeborová	Membre invité
Riccardo Zecchina	Rapporteur

Sommaire

Résumé

Un grand nombre de problèmes d'optimisation, ainsi que de problèmes inverses, combinatoires ou hors équilibre qui apparaissent en physique statistique des systèmes complexes, peuvent être représentés comme un ensemble de variables en interaction sur un certain réseau. Bien que la recette universelle pour traiter ces problèmes n'existe pas, la compréhension qualitative et quantitative des problèmes complexes sur des graphes a fait de grands progrès au cours de ces dernières années. Un rôle particulier a été joué par des concepts empruntés à la physique des verres de spin et la théorie des champs, qui ont eu beaucoup de succès en ce qui concerne la description des propriétés statistiques des systèmes complexes et le développement d'algorithmes efficaces pour des problèmes concrets.

En première partie de cette thèse, nous étudions des problèmes de diffusion sur des réseaux, avec la dynamique hors équilibre. En utilisant la méthode de cavité sur des trajectoires dans le temps, nous montrons comment dériver des équations dynamiques dites "message-passing" pour une large classe de modèles avec une dynamique unidirectionnelle – la propriété clef qui permet de résoudre le problème. Ces équations sont asymptotiquement exactes pour des graphes localement en arbre et représentent en général une bonne approximation pour des réseaux réels. Nous illustrons cette approche avec une application des équations dynamiques pour résoudre le problème inverse d'inférence de la source d'épidémie dans le modèle "susceptible-infected-recovered".

Dans la seconde partie du manuscrit, nous considérons un problème d'optimisation d'appariement planaire optimal sur une ligne. En exploitant des techniques de la théorie de champs et des arguments combinatoires, nous caractérisons une transition de phase topologique qui se produit dans un modèle désordonné simple, le modèle de Bernoulli. Visant une application à la physique des structures secondaires de l'ARN, nous discutons la relation entre la transition d'appariement parfait-imparfait et la transition de basse température connue entre les états fondu et vitreux de biopolymère ; nous proposons également des modèles généralisés qui suggèrent une correspondance exacte entre la matrice des contacts et la séquence des nucléotides, permettant ainsi de donner un sens à la notion des alphabets effectifs non-entiers.

1 Introduction : physique statistique des systèmes complexes

La physique statistique des systèmes complexes peut être vue comme un ensemble de concepts et de techniques qui permettent d'étudier en détail des systèmes composés d'un grand nombre de variables en interaction : particules, spins, couleurs, variables binaires, *etc.* Conceptuellement, le succès de la physique statistique réside en l'explication du comportement collectif émergent des systèmes à beaucoup de constituants à partir de leurs interactions élémentaires. Bien que les manifestations du comportement global peuvent être très différentes, elles peuvent être décrites par des concepts généraux, tels que la *transition de phase* – un changement qualitatif de l'état du système.

Il est souvent utile de représenter la topologie des interactions dans un système complexe par un *réseau* $G = (V, E)$, où V est l'ensemble des nœuds, correspondants aux variables, et E est l'ensemble des arêtes (dirigées ou non), correspondantes aux interactions entre elles. Ce réseau d'interaction peut être extrait à partir du système *réel*, ou bien peut être généré de façon artificielle ; dans ce dernier cas, des *graphes aléatoires* sont souvent utilisés comme une approximation du réseau d'interaction. Dans cette thèse, ces deux classes sont utilisées à titre égal pour une modélisation physique et pour une vérification de la performance des algorithmes dévéloppés. Dans la partie I, la plupart des réseaux seront dits *creux* et *localement en arbre*, définis dans un espace multi-dimensionnel, tandis que dans la partie II, nous allons nous concentrer sur des réseaux *denses* et *planaires*, c'est-à-dire définis sur un plan bidimensionnel.

L'évolution du système complexe est généralement accompagnée par une certaine dynamique microscopique. Néanmoins, en fonction de la question particulière qu'on pose sur les propriétés du système, le problème peut être classifié comme *statique* ou *dynamique* ; les deux classes peuvent être définies sur des graphes statiques ou dynamiques. Dans la première partie de cette thèse, une attention pariculière sera portée à l'étude de la dynamique des systèmes complexes *hors équilibre*. Dans la partie II, nous allons traiter un problème d'optimisation statique. Une des plus grandes difficultés qui apparaissent dans tous les problèmes considérés provient essentiellement de leur nature *désordonnée*, surtout dans le cas d'un désordre *gelé*.

2 Partie I. Processus dynamiques sur des graphes

La théorie des systèmes désordonnés à l'équilibre a fait des grands progrès au cours de ces dernières années. Les méthodes de champ moyen permettent de traiter des problèmes impliquant un grand nombre de variables en interaction, dans le cas des graphes denses ou dilués, ce qui a permis d'introduire des algorithmes efficaces pour l'inférence statistique et l'optimisation.

Malgré un considérable effort de construction des nouveaux outils analytiques afin de confronter une dynamique hétérogène et hors équilibre, une méthode efficace de résolution de la dynamique générale n'existe pas encore à ce jour. Dans cette partie, nous allons introduire une approche récente et prometteuse au problème, connue sous le nom de la méthode de cavité dynamique.

2.1 Exemple d'un processus dynamique

Avant de commencer à formuler une approche générale pour des problèmes dynamiques, prenons un exemple d'un processus dynamique concret et intuitif que nous allons utiliser dans la suite : le modèle "susceptible-infected-recovered" (SIR) qui est souvent employé pour décrire une propagation d'épidémie ou d'information. Dans ce modèle, un nœud dans le réseau peut être dans un des trois états : sain (S), infecté (I), ou guéri (R). La propagation d'infection dans ce modèle se passe via une interaction entre une paire d'individus dans les états I et S. À chaque pas de temps, un nœud infecté peut devenir guéri avec une certaine probabilité μ. Ces régles de transition peuvent être représentées par un schéma suivant :

$$S(i) + I(j) \xrightarrow{\lambda_{ji}} I(i) + I(j) \tag{1}$$

$$I(i) \xrightarrow{\mu_i} R(i). \tag{2}$$

Dans une première approximation, la dynamique (1)-(2) est étudiée sous une hypothèse de mélange homogène, qui consiste à négliger la topologie actuelle du réseau, et qui permet d'écrire des équations de champ moyen simples sur les densités des nœuds sains, infectés et guéris.

Quelques études plus récentes ont dévéloppé une approche plus avancée au problème, en écrivant des équations différentielles qui tiennent compte de l'environement hétérogène. Bien que moyennées sur les conditions initiales et des ensembles de graphes aléatoires, ces équations sont exactes dans la limite d'un grand nombre d'agents d'interaction, pour des graphes localement en arbre. Le calcul de ces équations est basé sur une identification des "justes" variables dynamiques qui sont nécessaires afin de pouvoir écrire des equations férmées.

Existe-il un moyen de calculer ce genre d'équations d'une façon plus systèmatique, et d'étendre cet analyse sur autres problèmes dynamiques ? Il se trouve que dans beaucoup de cas la réponse est positive ; la méthode que nous allons utiliser dans la suite est basée sur une approche appellée la méthode de cavité dynamique, décrite ci-dessous (pour plus de détails, consulter la publication [P-1]).

2.2 Méthode de cavité dynamique

Prenons un graphe $G = (V, E)$. Chaque nœud $i \in V$ est caractérisé par une variable qui prend une valeur σ_i^t à l'instant t. Nous supposons que l'ensemble des valeurs possibles de σ_i^t est de taille finie K. Supposons de plus qu'un processus dynamique générique est décrit par une probabilité de transition $w_i(\sigma_i^{t+1} \mid \{\sigma_j^t\}_{j\in\partial i})$ que le nœud i prend la valeur σ_i^{t+1} à l'instant $t+1$, étant données les valeurs $\{\sigma_j^t\}$ de ses voisins à l'instant t (l'ensemble des voisins de i sur le graphe est noté par ∂i).

Si on dénote $\vec{\sigma}_i = (\sigma_i^0, \ldots, \sigma_i^T)$ la trajectoire de i aux temps $t = 0, \ldots, T$, où T est défini comme un temps d'arrêt, la distribution de probabilié jointe $P(\{\vec{\sigma}_i\}_{i\in V})$ s'écrit :

$$P(\{\vec{\sigma}_i\}_{i\in V}) = \prod_{i\in V} \prod_{t=0}^{T-1} w_i(\sigma_i^{t+1} \mid \{\sigma_j^t\}_{j\in\partial i}) P_0, \tag{3}$$

où $P_0 \equiv P(\{\sigma_i^0\}_{i\in V})$ est une distribution des variables à l'instant initial.

Les équations de cavité (appelées également équations "belief propagation", ou BP, dans certains contextes) sont exactes sur des modèles graphiques quand le graphe des facteurs est un arbre. Pourtant, la dynamique définie par (3) transforme le graphe initial (même dans le cas où G est un arbre) en graphe avec un grand nombre de boucles, voir Fig. 1. Ce problème peut être contourné par le passage au graphe dual qui conserve la topologie initiale, comme montré sur la Fig. 2.

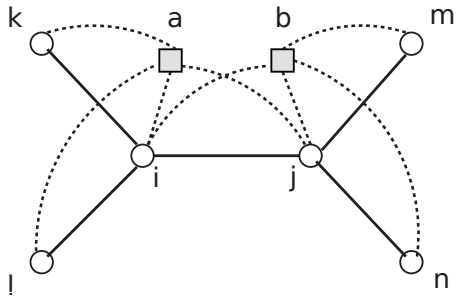

FIGURE 1 – Un exemple d'un graphe des facteurs pour le modèle graphique à deux temps adjacents, décrit par la distribution de probabilité $P(\{\vec{\sigma}_i\}_{i\in V})$. Le facteur a représente une interaction entre la variable σ_i^{t+1} et des variables $\{\sigma_j^t\}_{j\in\partial i}$ à l'instant précédent. Ce graphe des facteurs est caractérisé par l'apparition des boucles systematiques.

La probabilité jointe des trajectoires temporelles (3) peut être reécrite en termes de nouvelles variables auxiliaires $\vec{\sigma}_{i\to j}$, définies sur chaque arête dirigée $(i,j)\in E$:

$$P(\{\vec{\sigma}_{i\to j},\vec{\sigma}_{j\to i}\}_{(i,j)\in E}) = \prod_{i\in V}\prod_{t=0}^{T-1}\left[w_i(\sigma_{i\to l}^{t+1}\mid\{\sigma_{k\to i}^t\}_{k\in\partial i})\prod_{k\in\partial i\setminus l}\delta_{\sigma_{i\to l}^t,\sigma_{i\to k}^t}\right]P_0, \qquad (4)$$

où l est une des variables directement influencée par i.

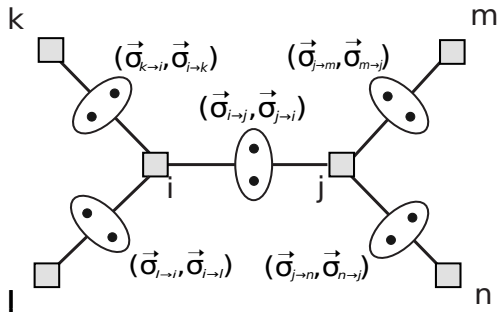

FIGURE 2 – Un exemple d'un graphe des facteurs pour le modèle graphique à tous temps, décrit par la distribution de probabilité $P(\{\vec{\sigma}_{i\to j},\vec{\sigma}_{j\to i}\}_{(i,j)\in E})$. Le facteur i représent une interaction entre les trajectoires $\vec{\sigma}_i$ et $\{\vec{\sigma}_j\}_{j\in\partial i}$. Ce graphe des facteurs est caractérisé par la structure topologique du graphe d'origine.

Cette propriété du graphe dual nous permet d'appliquer la méthode de cavité, et d'écrire des équations BP pour le modèle graphique défini par (4) :

$$m^{i\to j}(\vec{\sigma}_i,\vec{\sigma}_j) = \frac{1}{Z^{i\to j}}\sum_{\{\vec{\sigma}_k\}_{k\in\partial i\setminus j}}\left[\prod_{t=0}^{T-1}w_i(\sigma_i^{t+1}\mid\{\sigma_k^t\}_{k\in\partial i\setminus j},\sigma_j^t)P_0\right]\prod_{k\in\partial i\setminus j}m^{k\to i}(\vec{\sigma}_k,\vec{\sigma}_i), \qquad (5)$$

où on a choisi de rebaptiser les variables $\{\vec{\sigma}_{i\to j}, \vec{\sigma}_{j\to i}\}_{(i,j)\in E}$ en $\{\vec{\sigma}_i, \vec{\sigma}_j\}_{(i,j)\in E}$.

La constante de normalisation $Z^{i\to j}$ peut être calculée explicitement pour une dynamique Markovienne à partir de la condition

$$\sum_{\vec{\sigma}_i,\vec{\sigma}_j} m^{i\to j}(\vec{\sigma}_i, \vec{\sigma}_j) = 1, \tag{6}$$

ce qui donne

$$Z^{i\to j} = \frac{1}{2^{(T+1)(d_i-2)}} \tag{7}$$

dans ce cas, où d_i est le nombre des voisins de i dans le graphe initial.

Le message $m^{i\to j}(\vec{\sigma}_i, \vec{\sigma}_j)$ a le sens d'une probabilité des trajectoires $\vec{\sigma}_i, \vec{\sigma}_j$ dans le graphe de cavité transformé, où le facteur-nœud j a été enlevé. Les équations (5) peuvent également être reécrites en termes de probabilités conditionnelles $m^{i\to j}(\vec{\sigma}_i \mid \vec{\sigma}_j)$ dans le graphe de cavité : nous avons

$$\sum_{\vec{\sigma}_i} m^{i\to j}(\vec{\sigma}_i, \vec{\sigma}_j) = \frac{1}{2^{T+1}}, \tag{8}$$

et ainsi

$$m^{i\to j}(\vec{\sigma}_i \mid \vec{\sigma}_j) = \sum_{\{\vec{\sigma}_k\}_{k\in\partial i\setminus j}} \left[\prod_{t=0}^{T-1} w_i(\sigma_i^{t+1} \mid \{\sigma_k^t\}_{k\in\partial i\setminus j}, \sigma_j^t) P_0\right] \prod_{k\in\partial i\setminus j} m^{k\to i}(\vec{\sigma}_k \mid \vec{\sigma}_i). \tag{9}$$

Le facteur de normalisation est égal à 1 grace à la propriété Markovienne de la dynamique. Le message $m^{i\to j}(\vec{\sigma}_i \mid \vec{\sigma}_j)$ dans (9) a le sens d'une probabilité marginale pour une trajectoire $\vec{\sigma}_i$ étant donnée la trajectoire $\vec{\sigma}_j$ dans le graphe de cavité. L'équation (9) peut être résolue par itération jusqu'à une convergence ; la probabilité marginale pour la trajectoire $\vec{\sigma}_i$ dans le graphe complet sera ensuite donnée par

$$m^{i}(\vec{\sigma}_i) = \sum_{\{\vec{\sigma}_k\}_{k\in\partial i}} \left[\prod_{t=0}^{T-1} w_i(\sigma_i^{t+1} \mid \{\sigma_k^t\}_{k\in\partial i}) P_0\right] \prod_{k\in\partial i} m^{k\to i}(\vec{\sigma}_k \mid \vec{\sigma}_i). \tag{10}$$

Dans le cas général, la résolution de cette équation demande un nombre exponentielle d'opérations, puisque chaque message contient K^T composantes, ce qui veut dire que la somme dans (9) est prise sur $K^{T(d_i-1)}$ variables pour chaque nœud i, avec d_i le nombre de voisins de i. Toutefois, nous allons voir qu'une simplification cruciale apparaît pour des modèles avec une dynamique particulière, dite unidirectionnelle, introduits dans la section suivante.

2.3 Modèles avec une dynamique unidirectionnelle

Il s'avère qu'à partir des équations de cavité (9) et (10), il est possible de calculer des équations dynamiques de passage de messages (équations DMP) pour une grande classe de modèles à dynamique irréversible, c'est-à-dire définis par une probabilité de transmission $w_i(\sigma_i^{t+1} \mid \{\sigma_j^t\}_{j\in\partial i})$ où chaque valeur de σ_i^t peut prendre un des K états ordonnés de façon irréversible :

$$\Omega_1 \to \Omega_2 \to \ldots \to \Omega_K. \tag{11}$$

De nombreux exemples de ce genre de dynamique incluent la dynamique du modèle d'Ising dans un champ aléatoire à température nulle, la dynamique de propagation d'épidemies et de rumeurs, *etc.* Ici, nous allons illustrer cette approche avec les équations DMP pour le modèle SIR ; tous les détails concernants autres modèles peuvent être retrouvés dans la publication [P-1].

Nous commencons par remarquer que dans le modèle SIR, défini par des équations (1) et (2), la trajectoire d'un nœud i peut être complétement paramétrisée par seulement deux temps de changement d'état : $\vec{\sigma}_i = \left|S_0 SSSSSSI_{\tau_i} IIIIIIR_{\omega_i} RRRRRR_T\right\rangle \longleftrightarrow (\tau_i, \omega_i)$. Par conséquent, les probabilités de trouver le nœud i dans les trois états sont défini comme

$$P_S^i(t) = \sum_{\tau_i > t} \sum_{\omega_i > \tau_i} m^i(\tau_i, \omega_i), \tag{12}$$

$$P_I^i(t) = \sum_{\tau_i \leq t} \sum_{\omega_i > t} m^i(\tau_i, \omega_i), \tag{13}$$

$$P_R^i(t) = \sum_{\omega_i \leq t} \sum_{\tau_i < \omega_i} m^i(\tau_i, \omega_i). \tag{14}$$

Il sera également utile d'introduire une probabilité marginale dans un graphe de cavité :

$$P_S^{i\to j}(t) = \sum_{\tau_i > t} \sum_{\omega_i > \tau_i} m^{i\to j}(\tau_i, \omega_i \mid T, T). \tag{15}$$

Les messages ont les propriétés suivantes.

Propriété 1. $m^{i\to j}(\tau_i, \omega_i \mid T, T) = 0$ si $\tau_i \geq \omega_i$;

Propriété 2. Si $\tau_j \geq \tau_i$, alors $m^{i\to j}(\tau_i, \omega_i \mid \tau_j, \omega_j) = m^{i\to j}(\tau_i, \omega_i \mid t', \omega_j)$ pour chaque $\tau_i \leq t' < \omega_j$;

Propriété 3. $\sum_{\tau_i, \omega_i} m^{i\to j}(\tau_i, \omega_i \mid T, T) = 1$;

Propriété 4. $m^{i\to j}(\tau_i, \omega_i + 1 \mid T, T) = (1 - \mu_i) m^{i\to j}(\tau_i, \omega_i \mid T, T)$.

Des propriétés équivalentes à 1, 3 et 4 sont également valables pour les marginales $m^i(\tau_i, \omega_i)$. Il est facile d'établir les deux premières équations d'évolution sur des quantités $P_S^i(t)$, $P_I^i t$ et $P_R^i(t)$. Selon les définitions,

$$\begin{aligned} P_R^i(t+1) = \sum_{\omega_i \leq t+1} \sum_{\tau_i < \omega_i} m^i(\tau_i, \omega_i) &= \sum_{\omega_i \leq t} \sum_{\tau_i < \omega_i} m^i(\tau_i, \omega_i) + \delta_{\omega_i, t+1} \sum_{\tau_i \leq t} m^i(\tau_i, \omega_i) \\ &= P_R^i(t) + \mu_i P_I^i(t), \end{aligned} \tag{16}$$

où nous avons utilisé un équivalent de la propriété 4 pour les marginales, puisque

$$\sum_{\omega_i \geq t+1} m^i(\tau_i, \omega_i) = \frac{1}{1 - (1 - \mu_i)} m^i(\tau_i, t+1) = \frac{1}{\mu_i} m^i(\tau_i, t+1). \tag{17}$$

Comme la somme des expressions définies en (12)-(14) donne un, il est évident que

$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1). \tag{18}$$

Après quelques manipulations algébriques (les détails des caluls sont donnés en [P-1]), on peut montrer que $P_S^{i\to j}(t+1)$ peut être mis sous une forme

$$P_S^{i\to j}(t+1) = P_S^i(0) \prod_{k\in\partial i\setminus j} \theta^{k\to i}(t+1), \tag{19}$$

où à chaque pas de temps $\theta^{k\to i}(t+1)$ est calculé à partir de $P_S^{k\to i}(t)$ comme suit :

$$\theta^{k\to i}(t+1) - \theta^{k\to i}(t) = -\lambda_{ki}\phi^{k\to i}(t), \tag{20}$$
$$\phi^{k\to i}(t) = (1-\lambda_{ki})(1-\mu_k)\phi^{k\to i}(t-1) - [P_S^{k\to i}(t) - P_S^{k\to i}(t-1)]. \tag{21}$$

Les probabilités marginales que i soit dans un état donné à l'instant t sont alors données par

$$P_S^i(t+1) = P_S^i(0) \prod_{k\in\partial i} \theta^{k\to i}(t+1)\,, \tag{22}$$
$$P_R^i(t+1) = P_R^i(t) + \mu_i P_I^i(t)\,, \tag{23}$$
$$P_I^i(t+1) = 1 - P_S^i(t+1) - P_R^i(t+1)\,, \tag{24}$$

avec des conditions initiales pour des messages dynamiques :

$$\theta^{i\to j}(0) = 1, \tag{25}$$
$$\phi^{i\to j}(0) = \delta_{\sigma_i^0, I}. \tag{26}$$

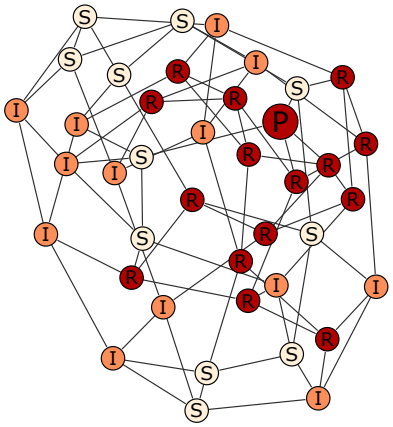

FIGURE 3 – Un exemple d'une instance donnée du problème d'inférence sur un graphe régulier aléatoire de degré $c = 4$ avec $N = 40$ nœuds. Le patient zéro est indiqué par l'étiquette P et apparaît dans un état R dans l'observation. L'épidémie est générée avec des paramètres $\lambda = 0.5$ et $\mu = 0.5$, l'obervation est présentée à l'instant $t_0 = 5$.

Ces équations DMP sont asymptotiquement exactes sur des graphes localement en arbre et peuvent être appliquées à une instance donnee du graphe. La section suivante donne une illustration d'application de ces équations, où ces propriétés sont cruciales pour la résolution du problème.

2.4 Estimation de la source d'épidémie avec les équations DMP

Dans cette section, nous allons présenter une application des équations (16)-(26) pour le modèle SIR, à un problème inverse de détéction de l'origine de l'épidémie, étant donné l'état du réseau à un certain moment ; les détails peuvent être trouvés dans l'article [P-2].

Le problème est défini comme suit. Supposons qu'à l'instant $t = 0$, uniquement un nœud est infecté (on va appeler ce nœud le *patient zéro*), et tous les autres nœuds sont dans l'état S. Après $t_0 > 0$ pas de temps (t_0 est en général inconnu), l'état $\mathcal{O}$ du système est observé, et l'objectif est d'estimer le patient zéro à partir de cet état (voir Fig. 3).

L'algorithme proposé est basé sur les équations DMP qui permettent d'estimer les probabilités $P_S^j(t, i_0)$ (respectivement $P_I^j(t, i_0)$, $P_R^j(t, i_0)$) que le nœud j est dans un de trois états S, I, ou R, à l'instant t, pour un patient zéro i_0 donné. Pour le moment, imaginons que le temps d'observation est connu. Avec la règle de Bayes, la probabilité que i est le patient zéro $P(i|\mathcal{O}) \sim P(\mathcal{O}|i)$. Si on savait calculer $P(\mathcal{O}|i)$ de façon exacte, le nœud pour lequel "l'énergie" $E(i) \equiv -\log P(\mathcal{O}|i)$ est maximisée sera la solution optimale du problème. Malheureusement, on peut seulement espérer calculer la probabilité $P(\mathcal{O}|i)$ approximativement ; une possiblité sera de l'approcher par le produit des probabilités marginales qui proviennent des équations DMP :

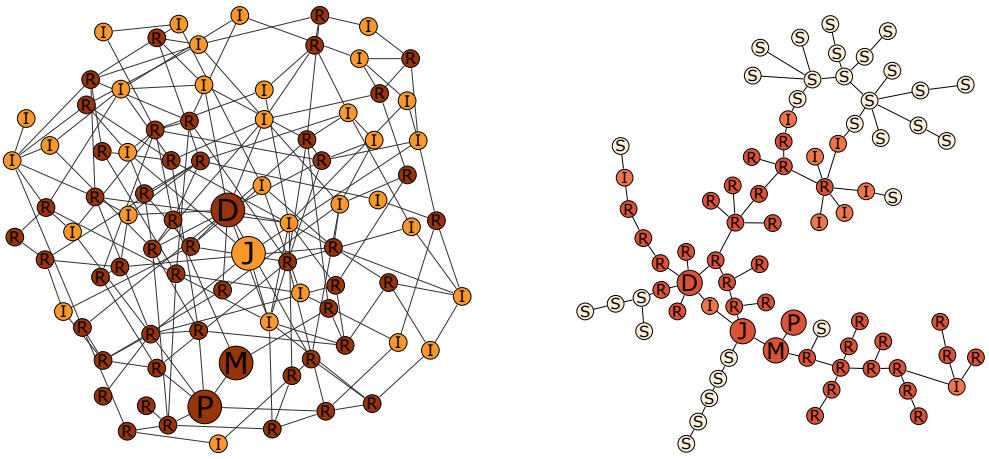

FIGURE 4 – Gauche : une instance du problème d'inférence sur un graphe Erdös-Rényi avec un dégré moyen $\langle c \rangle = 4$ et $N = 84$. L'épidemie est générée avec des paramètres $\lambda = 0.7$ et $\mu = 0.5$. Dans cet exemple, uniquement des nœud infectés (clair) et geurris (sombre) sont présents à l'instant $t_0 = 5$. Droite : une instance du problème d'inférence sur un graphe "scale-free" avec un dégré moyen $\langle c \rangle = 5/3$ et $N = 77$. L'épidemie est générée avec des paramètres $\lambda = 0.7$ et $\mu = 0.5$, l'obervation est présentée à l'instant $t_0 = 10$. Dans les deux figures, le vrai patient zéro est étiqueté par P, les meilleurs estimations par les méthodes de DMP, ainsi que des centralités basées sur l'estimateur de Jordan et de distance, sont marquées par M, J et D, respectivement.

$$P(\mathcal{O}|i) \simeq \prod_{\substack{k \in \mathcal{O} \\ \sigma_k(t_0)=S}} P_S^k(t,i) \prod_{\substack{l \in \mathcal{O} \\ \sigma_l(t_0)=I}} P_I^l(t,i) \prod_{\substack{n \in \mathcal{O} \\ \sigma_n(t_0)=R}} P_R^n(t,i). \tag{27}$$

Pour estimer la valeur de t_0, nous pouvons calculer l'énergie $E(i,t)$ pour différentes valeurs de t, et choisir la valeur qui maximise la "fonction de partition" $Z(t) \equiv \sum_i e^{-E(i,t)}$. La complexité algorithmique pour le calcul de $E(i)$ pour un vertex i donné s'elève à $\mathcal{O}(t_0 N c)$, où c est un dégré moyen dans le graphe.

En [P-2], la performance de cette méthode est comparée aux algorithmes basés sur des mesures de centralité dans les graphes (cf. Fig. 4) ; l'algorithme DMP donne des meilleurs résultats dans la grande majorité des cas.

3 Partie II. Appariement planaire et repliement optimal

Optimisation est un concept omniprésent dans la nature en général, et dans les activités humaines en particulier. Un grand nombre de problèmes très variés peuvent être formulés comme des probèmes d'optimisation, qui consistent à trouver une configuration des variables minimisant une certaine fonction de coût.

Des *problèmes d'appariement* sur des graphes sont souvent utilisés pour tester de nouvelles idées de la science de la complexité. En général, la recherche d'un *appariement maximal* est un problème relativement facile d'une *complexité polynomiale*, tandis que le calcul d'un *nombre d'appariements* est classifié comme un problème *exponentiellement difficile* dans la classe NP. Dans cette partie, nous allons présenter un problème d'appariement particulier sous une contrainte globale de planarité, dans lequel les deux objectifs peuvent être atteints en nombre polynomial d'opérations. Malgré cette simplicité algorithmique, l'analyse analytique de ce problème reste assez compliqué. Un intérêt particulier pour le problème d'appariement planaire vient de sa pertinence pour la mécanique statistique des *structures sécondaires* de l'ARN.

3.1 Modèle de Bernoulli

Dans cette section, nous allons définir un modèle très simple d'appariement planaire désordonné. Prenons L points $i = 1, \ldots, L$ sur une ligne droite, et définissons une matrice symétrique des contacts possibles A comme une matrice aléatoire avec des éléments distribués indépendamment selon la loi

$$\mathrm{Prob}(A_{ij}) = p\delta(A_{ij} - 1) + (1 - p)\delta(A_{ij}), \tag{28}$$

où $\delta(x) = 1$ pour $x = 0$, et $\delta(x) = 0$ sinon. En d'autres termes, chaque élément $A_{ij} = A_{ji}$ est indépendamment égal à un avec probabilité p pour chaque $i \neq j$, ou à zéro sinon. On dessine $L/2$ arches non-croisées entre les paires des points, autorisés par les valeurs non-nulles de A_{ij}, de façon à ce que chaque point participe dans une connexion exactement, voir Fig. 5(a). Si au moins un ensemble de tels liens existe, nous allons dire que le problème a une solution d'*appariement parfait*.

Une représentation importante des diagrammes planaires à L points est donné par des chemins de Dyck, voir Fig. 5. Le nombre total des chemins de Dyck avec un nombre pair des pas est donné par le nombre de Catalan

$$C_{L/2} = \frac{L!}{(\frac{L}{2})!(\frac{L}{2}+1)!} \sim \frac{2^L}{L^{3/2}}\sqrt{\frac{2^3}{\pi}}, \tag{29}$$

où l'expression asymptotique est valable pour $L \gg 1$.

Quand $0 < p < 1$, certains diagrammes ne sont pas autorisés. Ceci réduit le nombre de solutions au problème d'appariement maximal, qui devient zéro au-dessous d'une certaine

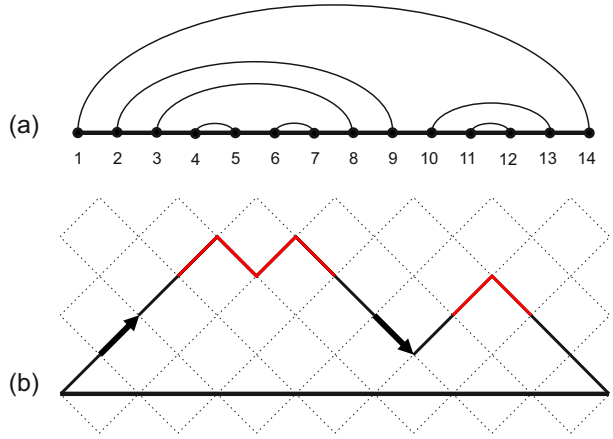

FIGURE 5 – Un exemple de (a) configuration d'appariement planaire parfait, et (b) représentation comme un chemin de Dyck. Une arche est donnée par les pas "vers le haut" et "vers le bas" à la même hauteur, presentés par les flèches ↗ et ↘. Une partie du chemin entre les flèches est également un chemin de Dyck. Les arches de longueur minimale correspondent à des pics dans la représentation de chemin de Dyck (marqués en rouge).

probabilité critique p_c. Cette probabilité critique peut être determinée numériquement avec un algorithme itératif suivant (pour plus de détails, cf. publication [P-3]) :

$$F_{i,i+k} = \lim_{T\to 0} T \ln Z_{i,i+k} = \max_{s=i+1,\ldots,i+k} \left[F_{i+1,i+k}, \varepsilon_{i,s} + F_{i+1,s-1} + F_{s+1,i+k}\right]. \qquad (30)$$

Cette expression peut être interpétée comme une croissance d'un graphe optimal : à chaque pas de temps, un nœud est ajouté à la séquence, et les liens dans le nouveau graphe sont redistribués pour minimiser le nombre de lacunes, cf. Fig. 6. Cet algorithme dit de programmation dynamique a une complexité cubique en longueur L.

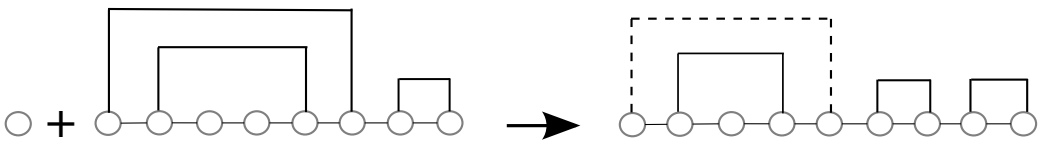

FIGURE 6 – Une inteprétation de l'algorithme (30) en termes d'un graphe croissant.

L'algorithme exact (30) permet de détecter le point $p_c \simeq 0.379$ de transition de phase dans l'étude de la probabilité de trouver au moins une solution au problème d'appariement maximal : les graphiques correspondants sont présentés sur la Fig. 7(a) pour des longueurs de polymères différentes, $L = 500, 1000, 2000$. L'analyse d'échelle est presenté sur la Fig. 7(b).

Il se trouve que le calcul analytique de la valeur critique p_c est un problème difficile à cause de la nature gélée du désordre dans le problème [P-3]. Dans la publication [P-4], une méthode combinatoire itérative d'expansion en longueurs des arches est developpée, permettant d'arriver à une estimation $p_c^* \simeq 0.3743$ au deuxième ordre d'expansion. Un intérêt supplémentaire de l'étude de cette transition provient de sa connexion avec la physique des structures sécondaires de l'ARN à basse température, expliquée dans la section suivante.

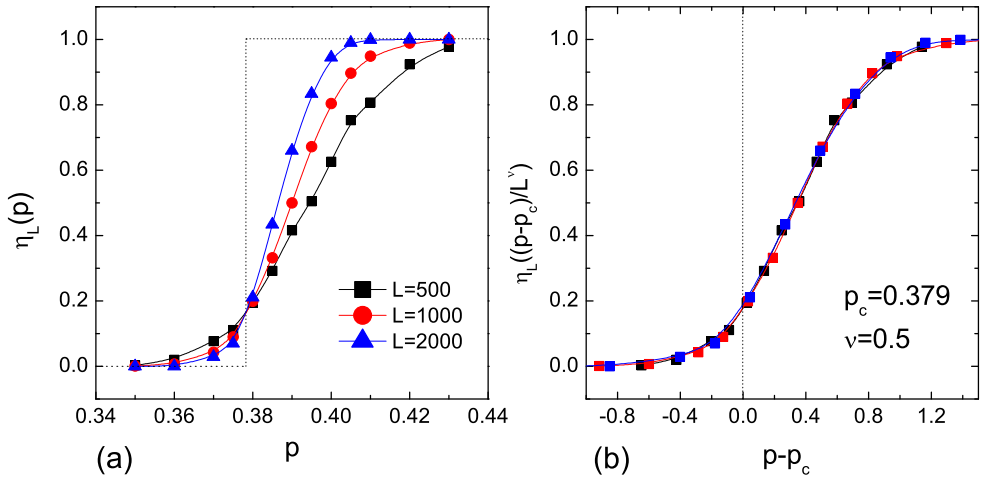

FIGURE 7 – (a) La fraction des appariements parfaits $\eta_L(p)$ comme une fonction de la densité p d'uns dans la matrice de contacts A pour des longueurs de la chaine $L = 500, 1000, 2000$, moyennées sur 10000 instances. La ligne pointillée correspond à la limite thermodynamique $L \to \infty$, ce qui donne la valeur critique $p_c = 0.379$. (b) L'analyse d'échelle des lignes, correspondants aux différentes longueurs de la chaine L. La procédure de fitting donne l'exposant de la largeur de distribution $\nu \approx 0.5$.

3.2 Mécanique statistique des structures sécondaires de l'ARN

L'ARN réel est un biopolymère composé d'une séquence de quatre aminobases, avec des règles fixes d'appariement. Dans un état replié de l'ARN, des structures "cactus" (dessinées sur les Fig. 8(a) et Fig. 8(c)) sont cinétiquement favorisées, tandis que des configurations contenant des pesudonœuds sont défavorisées.

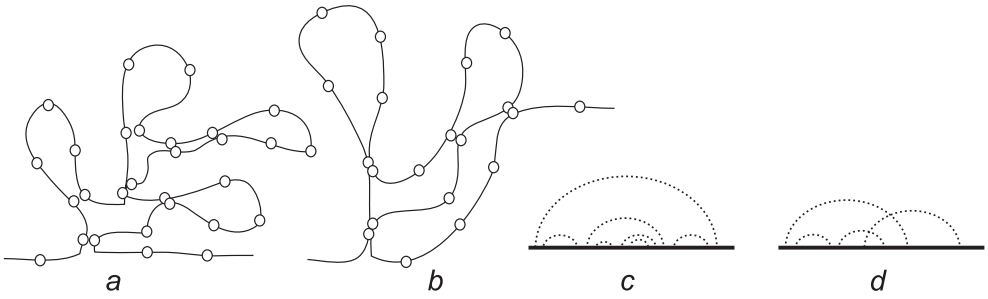

FIGURE 8 – Exemples schématiques des structures sécondaires de l'ARN et ses représentations de contacts corréspondantes : avec [(b) et (d)] et sans [(a) et (c)] les pesudonœuds.

La transition de phase presentée dans la section précédente dans le contexte de physique de l'ARN correspondrait à l'optimalité de la structure (à température zéro) en fonction de la varation de l'alphabet effectif. Dans la publication [P-3], un diagramme de phase complet du modèle de Bernoulli, ainsi que le lien avec une transition "fondue-gelée" à une température non-nulle ont été considérés, résultant en une conjecture que le point de transition entre les structures d'appariement complet et non-complet correspond à un point terminal de la ligne critique des températures de la transition "fondue-gelée", voir Fig. 9 ; pour plus de détails, consulter la publication [P-3].

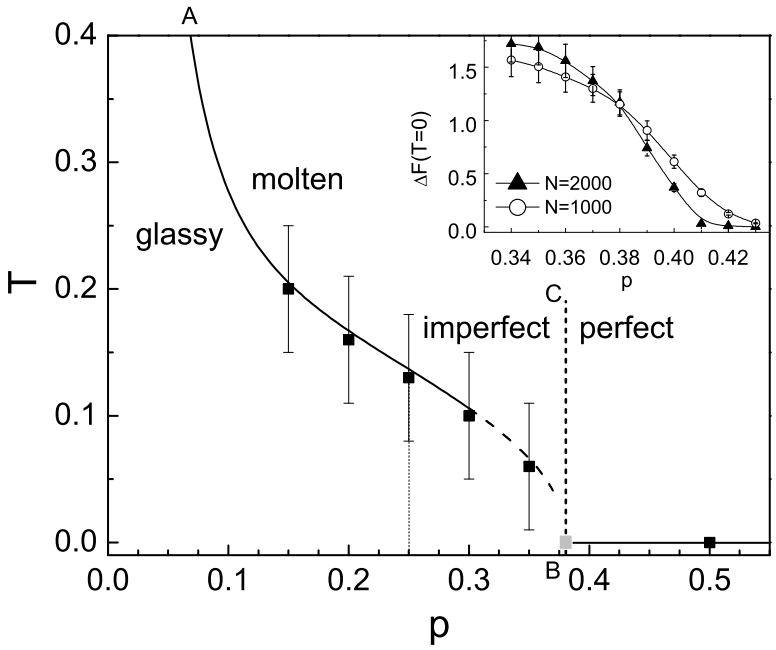

FIGURE 9 – Figure principale : le diagramme de phase du modèle de Bernouilli sur le plan (T, p). Les points correspondent aux températures critiques T_c de la transition "fondue-gelée" pour les différentes valeurs de $p = 0.15, 0.2, 0.25, 0.3, 0.35, 0.5$. La ligne critique (A-B) sépare les phases fondue et gelée. Nous présumons que le point B de la ligne coïncide avec le point critique p_c. Figure en haut à droite : un argument supplémentaire en faveur de cette conjecture. Étude de "l'énergie libre de pinçade" en fonction du paramètre p (pour plus de détails, voir [P-3]).

Liste des publications

[P-1] Andrey Y. Lokhov, M. Mézard, and Lenka Zdeborová. Dynamic message-passing equations for models with unidirectional dynamics. *Phys. Rev. E* **91**, 012811, 2015.

[P-2] Andrey Y. Lokhov, M. Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E* **90**, 012801, 2014.

[P-3] Andrey Y. Lokhov, Olga V. Valba, Mikhail V. Tamm, and Sergei K. Nechaev. Phase transition in random planar diagrams and RNA-type matching. *Phys. Rev. E* **88**, 052117, 2013.

[P-4] Andrey Y. Lokhov, Olga V. Valba, Sergei K. Nechaev, and Mikhail V. Tamm. Topological transition in disordered planar matching : combinatorial arcs expansion. *J. Stat. Mech*, P12004, 2014.