Modèle p-spin sur des hypergraphes aléatoires : des systèmes vitreux aux ensembles aléatoires d'équations linéaires booléennes

Renaud Detcherry & David Jarossay

Mémoire de maîtrise

Sous la direction de Marc Lelarge et Guilhem Semerjian

7 septembre 2009

Introduction

Un système aléatoire d'équations linéaires booléennes à N variables est la donnée d'un ensemble d'équations linéaires booléeennes $(E_i)_{i\in I}$, et de variables aléatoires $(N_i)_{i\in I}$, égales au nombre de copies de l'équation E_i dans le système.
Nous étudions dans cet exposé le cas où $\{E_i, i \in I\}$ est l'ensemble des équations faisant intervenir un nombre fixé p de variables distinctes parmi un nombre N de variables $x_1...x_N$ et où les N_i sont des variables de Bernoulli indépendantes de paramètre $\frac{\alpha}{N^{p-1}}$ où α est un paramètre. Ainsi, on a $|I| = 2\binom{N}{p}$

Un système d'équations linéaires booléennes est dit satisfiable s'il admet une solution. L'étude de la satisfiabilité du système précédent constitue le problème p-XORSAT. Plus précisément, nous étudions la probabilité $P_{SAT}(\alpha, N)$ pour que le système soit satisfiable en fonction du paramètre α et du nombre de variables N, on s'intéresse particulièrement à la limite thermodynamique $N \to \infty$ et à la probabilité $P_{SAT}(\alpha) = \lim_{N\to\infty} P_{SAT}(\alpha, N)$. Cette probabilité est décroissante en fonction de α, on s'intéresse alors à ses discontinuités. On s'intéressera enfin aux propriétés de l'ensemble des solutions lorsque le système est satisfiable.

Le phènomène remarquable est que P_{SAT} subit une transition de phase lorsque α franchit un seuil critique α_c, i.e. présente en ce point une discontinuité (si $p \geq 3$) ou une singularité (si p=2). Cette transition s'interprète physiquement ; il est en effet possible de relier le problème de satisfiabilité à des modèles physiques pour lesquels les transitions de phase sont bien connues.
Notons x_1,...,x_N les variables du systèmes, et posons $\sigma_i = (-1)^{x_i}$ pour tout i (σ_i représente physiquement un spin d'Ising). A une équation $x_{i_1} + ... + x_{i_p} = y$, on associe l'énergie de couplage entre p spins $\frac{1}{2}(1 - (-1)^y \sigma_{i_1}...\sigma_{i_p})$, qui vaut 0 si l'équation est satisfaite, et 1 sinon. L'énergie totale H du système, égale à la somme des énergies associées à chaque équation, est alors égale au nombre d'équations non satisfaites. En particulier, lorsque le système est satisfiable, les solutions du système correspondent aux états d'énergie minimale. Dans le cas d'un système homogène, les couplages entre spins sont tous les mêmes, on reconnaît un modèle d'Ising. Pour un système non homogène, il s'agit de modèles de verres de spin.
Nous étudions la transition de phase d'abord dans 2-XORSAT, puis dans p-XORSAT pour $p \geq 3$.

L'étude de l'existence d'états d'énergie nulle correspond, vis à vis des systèmes physiques modèlisés par ces modèles, et décrits par le hamiltonien précédent, à une étude à température nulle. Nous terminerons par une étude de ces systèmes à température non nulle, en s'intéressant notamment à la fonction de partition canonique et à l'énergie libre associée au hamiltonien $\mathcal{H}$.

Table des matières

1 2-XORSAT

1.1 Présentation des phénomènes

Dans le cas 2-XORSAT, les équations sont de la forme $x_i + x_j = y$. On associe alors au système un graphe aléatoire : une variable x_i correspond à un sommet, et une équation $x_i + x_j = y$ correspond à une arête reliant x_i et x_j étiquetée par y.

Intuitivement, un graphe fortement connecté correspondra à un système ayant peu de chances d'être satisfiable, et inversement, un graphe peu connecté correspondra à un système ayant beaucoup de chances d'être satisfiable.
De fait, la transition de satisfiabilité se produit au moment où le graphe subit une transition dite de percolation, c'est-à-dire que la taille maximale d'une composante connexe du graphe devient extensive. Physiquement, l'existence d'une composante connexe géante se traduit par des corrélations à longue distance entre les spins. Avant d'étudier la satisfiabilité dans 2-XORSAT, présentons donc ces phénomènes de percolation.

1.2 Percolation dans les graphes aléatoires

Pour le système aléatoire décrit dans l'introduction, décrivons la loi du graphe aléatoire. La présence de certaines arêtes dans le graphe sont des évènements indépendants, puisque la présence dans le système d'équations faisant intervenir des couples de variables différents sont des évènements indépendants. De plus, chacune des deux équations faisant intervenir deux variables données est présente avec probabilité $\frac{\alpha}{N}$, donc une arête donnée est présente dans le graphe avec probabilité $1 - (1 - \frac{\alpha}{N})^2 \sim \frac{\lambda}{N}$ où $\lambda = 2\alpha$.

Un graphe d'Erdös-Renyi $G(N, p)$ est un graphe aléatoire, à N sommets, pour lequel la présence de chaque arête a une probabilité p, et ces évènements sont tous indépendants. Le graphe associé au système est donc un graphe d'Erdös-Renyi. Dans un tel graphe, le degré d'un sommet donné a pour loi Bin(N,p). Pour que le degré moyen d'un sommet reste borné lorsque N varie, il est courant de supposer $p = \frac{\lambda}{N}$ où λ est un paramètre. Nous présenterons dans cette section un résultat de transition de phase pour la taille des composantes connexes de tels graphes en fonction du paramètre λ. On se rendra compte que seul le terme dominant dans p influe sur la transition de phase ; on pourra donc appliquer le même résultat de transition de phase au graphe associé au système, pour lequel $p = 1-(1-\frac{\alpha}{N})^2 \sim \frac{\lambda}{N}$.

Théorème 1.1. On peut distinguer deux régimes :
- Le régime sous-critique $\lambda < 1$. Les composantes connexes sont toutes de petite taille : si |C| désigne la taille de la plus grande composante connexe, on a l'estimation suivante :

$$\exists a \ / \ \mathbb{P}(|C| > a\, log(N)) \longrightarrow_{N\to\infty} 0$$

- Le régime sur-critique $\lambda > 1$. On voit l'apparition d'une composante connexe géante, si $|C_x|$ est la taille de la composante connexe d'un sommet, alors :

$$\exists \delta > 0 \ / \ \mathbb{P}(max(|C_x|) < \delta N) \longrightarrow_{N \to \infty} 0$$

Preuve. Pour prouver ces deux résultats, on cherche à estimer la taille d'une composante connexe à l'aide d'une marche aléatoire grâce à l'algorithme suivant qui génère la composante connexe C d'un sommet :

1) On se fixe un sommet x_0 du graphe, on pose $S_0 = 1$ et $A_0 = \{x_0\}$
2) On choisit aléatoirement un élément x de A_j. On tire aléatoirement le nombre de sommets auquel ce sommet est connecté dans le reste du graphe. Dans le cas d'Erdös-Rényi qu'on étudie, c'est une variable aléatoire de loi $Bin(n-k, \frac{\lambda}{N})$ où k est le nombre de sommets déjà reconnus comme dans la composante connexe de x ($k = |\cup_{i=0}^{j} A_i|$). On pose $A_{j+1} = A_j \cup \{y \in \{1...N\} \setminus \cup_{i=0}^{j} A_i / \text{ l'arête (x,y) est dans le graphe}\} - \{x\}$ et $S_j = |A_j|$
3) On recommence à l'étape 2), si A_{j+1} est non vide.

Au cours de l'algorithme, A est l'ensemble des sommets du graphe dits "actifs" c'est-à dire ceux qui restent à prendre en compte pour déterminer C_x, et $S_j = |A_j|$. Les autres sommets qui ont été dans A sont dits "explorés". De plus, l'algorithme introduit implicitement le temps d'arrêt $T = \inf\{j / S_j = 0\}$, T est la taille de la composante connexe de x.
Remarquons que les incréments de S sont indépendants, car ces évènements font toujours intervenir des ensembles d'arêtes disjoints. D'autre part, on voit que l'on peut majorer les incréments de S par des variables $-1 + Bin(N, \frac{\lambda}{N})$. Ces considérations permettent de prouver simplement le résultat pour $\lambda < 1$:
Soit $\lambda < 1$, soit R_j une marche aléatoire d'incréments indépendants $X_i = -1 + Bin(N, \frac{\lambda}{N})$ issue de $S_0 = 0$ et $T = \inf\{j / R_j = 0\}$. Par la loi faible des grands nombres, $T < +\infty$ presque sûrement. On a

$$\mathbb{E}(exp(\theta X_i)) = e^{-\theta}(1 - \frac{\lambda}{N} + (\frac{\lambda}{N})e^{\theta})^N \leq exp(-\theta + \lambda(e^{\theta} - 1)) = \phi(\theta) \qquad (1)$$

La fonction $\phi(\theta)$ est minimale en $\theta_1 = -\log(\lambda)$ et $\phi(\theta_1) < 1$. On introduit donc la sur-martingale positive $M_j = \frac{\exp(\theta_1 * R_j)}{\phi(\theta_1)^j}$. Par le lemme de Fatou et la convergence presque sûre des martingales positives on a $\mathbb{E}(\phi(\theta_1)^{-T}) \leq e^{\theta_1} = \frac{1}{\lambda}$ d'où par l'inégalité de Markov : $\mathbb{P}(T \geq k) \leq \frac{e^{klog(\phi(\theta_1))}}{\lambda}$. En prenant $k = \frac{(1+\varepsilon)(\log(N)}{-\log(\phi(\theta_1))}$, si $|C_x|$ désigne la taille de la composante connexe de x, on a

$$\mathbb{P}(|C_x| \geq \frac{(1+\varepsilon)(\log(N))}{-\log(\phi(\theta_1))}) \leq \frac{N^{-(1+\varepsilon)}}{\lambda} \qquad (2)$$

En sommant sur tous les sommets, on obtient le résultat annoncé.

Pour minorer la taille de $|C_x|$ dans le régime surcritique, nous avons besoin de modifier la marche aléatoire précédente. Soit δ tel que $(1-\delta)\lambda > 1$. On introduit une marche aléatoire W_j avec $W_0 = 1$ et d'incréments indépendants de loi $-1 + Bin((1-\delta)N, \frac{\lambda}{N})$. On voit que dans l'algorithme précédent pour générer la composante connexe de x, W_j minore en loi la taille de A_j tant que le nombre de sommets explorés reste inférieur à δN. Introduisons de nouveau un temps d'arrêt $T = \inf\{j/W_j = 0\}$. Par la remarque précédente on a $\mathbb{P}(|C_x| < \delta N) < \mathbb{P}(T < \delta N)$. A nouveau, pour majorer cette probabilité, introduisons une surmatingale adaptée, pour $\theta \in \mathbb{R}$ on a :

$$\mathbb{E}(\exp(\theta(-1 + Bin((1-\delta)N, \frac{\lambda}{N})) \leq \exp(-\theta + \lambda(1-\delta)(e^\theta - 1) = \psi_\delta(\theta) \tag{3}$$

Remarquons que $\lim_{\theta \to -\infty} \psi_\delta(\theta) = +\infty$ et $\psi_\delta^{'}(0) = (1-\delta)\lambda - 1 > 0$, donc il existe $\theta_\delta > 0$ tel que $\psi_\delta(-\theta_\delta) = 1$. $M_j = \exp(-\theta_\delta W_j)$ est une surmartingale positive, en stoppant au temps d'arrêt $T \wedge \delta N$, nous obtenons :

$$e^{-\theta_\delta} \geq \mathbb{E}(\exp(-\theta_\delta W_{T \wedge \delta N})) \geq \mathbb{P}(T \leq \delta N) \tag{4}$$

Nous avons donc prouvé qu'avec probabilité minorée par une constante $1 - e^{-\theta_\delta} > 0$, la composante connexe de x est de taille plus grande que δN.

Ici comme dans toute la suite de l'exposé, on dit qu'une propriété (P) pour un espace de proba $(\Omega_N, \mathbb{P}_N)$ a lieu asymptotiquement presque sûrement (et nous abrégerons par a.p.s.) si la probabilité de (P) tend vers 1 pour N tendant vers l'infini. Nous cherchons donc à montrer que la plus grande composante connexe est asymptotiquement presque sûrement de taille $> \delta N$, il faut pour cela modifier légèrement le raisonnement. Remarquons que dans les calculs précédents, si on avait eu $W_0 = m$ on aurait obtenu $\mathbb{P}(T \leq \delta N) \leq e^{-m\theta_\delta}$. Pour obtenir avec forte probabilité une composante connexe géante, il faut partir d'un sommet de fort degré.
Prenons $a > 0$ et $\delta^{'} > 0$ de sorte que $(1-a)(1-\delta^{'}) = (1-\delta)$. Isolons aN sommets du reste du graphe. Chacun a un nombre de voisins dans la partie à (1-a)N sommets de loi $Bin((1-a)N, \frac{\lambda}{N}))$ qui tend vers une loi de Poisson $(1-a)\lambda$. Donc pour $N \longrightarrow \infty$ chacun a une probabilité minorée par une constante > 0 d'avoir au moins m voisins dans l'autre partie du graphe, et donc avec probabilité qui tend vers 1 il y en a au moins un parmi les aN qui a plus de m voisins. On prend un tel sommet comme sommet initial dans l'algorithme, on a $|A_1| \geq m$. Si on continue l'algorithme dans le graphe à (1-a)N sommets, comme les arêtes internes de ce sous graphes sont indépendantes des degrés des aN premiers sommets, on obtient donc

$$\liminf_{N \to \infty} \mathbb{P}(\max(|C_x|) \leq \delta N) \leq e^{m\theta_{\delta'}} \ \forall m \in \mathbb{N}$$

□

Pour finir, notons qu'il existe des résultats beaucoup plus précis sur la taille de la compo-

sante connexe géante dans le régime surcritique. Nous énonçons, sans le démontrer car il n'est pas utile pour étudier la transition de satisfiabilité, le résultat suivant :

Théorème 1.1-bis. Soit $\lambda > 1$. Soit ρ, l'unique solution < 1 de l'équation $\rho = \exp(\lambda(\rho - 1))$. Alors il existe $\beta > 0$ tel que, asymptotiquement presque sûrement, dans le graphe aléatoire d'Erdös-Rényi de paramètre λ, il n'y ait qu'une seule composante connexe de taille au moins $\beta \log(N)$. De plus, la proportion de sommets du graphe qui sont dans cette composante tend en loi vers $1 - \rho$.

On pourra trouver dans [1] au chapitre 2 une preuve détaillée du théorème 1.1-bis, ainsi qu'une étude très complète des graphes d'Erdös-Rényi et de leur transition de percolation.

Visualisons le comportement de ρ en fonction de λ :

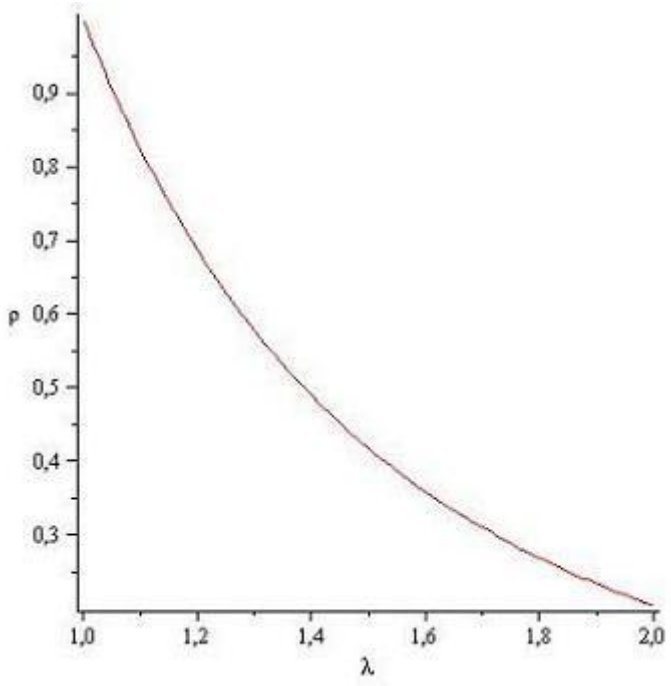

FIGURE 1 – Comportement de ρ en fonction de $\lambda = 2\alpha$

On observe que $\rho \to 1$ lorsque $\lambda \to 1^+$. Avec la dernière assertion du théorème ci-dessus, on voit que la proportion de sommets dans la composante connexe géante est continue au seuil de transition.

1.3 Transition de satisfiabilité

Cette partie vise à calculer $P_{SAT}(\alpha)$ à partir d'une estimation du nombre de cycles dans le graphe associé au système (la présence de cycles joue, on va l'expliquer dans ce qui suit, un grand rôle vis à vis de la satisfiabilité du système). La méthode du calcul de $P_{SAT}(\alpha)$ est issue de [2].

1.3.1 Calcul de P_{SAT} lorsque $\alpha > \frac{1}{2}$

Les résultats précédents nous permettent de calculer P_{SAT} pour $\alpha > \frac{1}{2}$. On observe que ce qui gêne la satisfiabilité du système est la présence de cycles dans le graphe. S'il n'y a aucun cycle, alors le système est toujours satisfiable : le graphe associé prend la forme d'un arbre (ou d'une union disjointe d'arbres), et on peut trouver une solution en attribuant une valeur 0 ou 1 à la variable "racine", puis les étiquettes des arêtes (on rappelle que dans le graphe aléatoire associé au système, les arêtes sont étiquetées par le second membre de l'équation auquelles elles correspondent) imposent les valeurs de ses "fils" dans le graphe, puis de tous ses descendants. Chaque cycle introduit une incompatibilité avec probabilité $1/2$: en effet, un cycle est satisfiable si et seulement si il a un nombre pair d'arêtes étiquetées par 1 : chaque étiquette 1 dans le cycle impose un changement de valeur entre une variable x_1 et une variable x_2 voisine dans le cycle. Si un cycle est satisfiable, alors on doit retrouver la même valeur pour x_1 après un tour de cycle, il doit donc y avoir un nombre pair d'arêtes étiquettées par 1. Réciproquement si ce nombre est pair, on attribue une valeur 0 ou 1 à une variable x_1 du cycle, on parcourt le cycle en attribuant des valeurs aux variables du cycle : la même que pour la variable précédente si l'arête les joignant est étiquettée 0, une valeur différente sinon. On obtient alors une solution. Le nombre d'arêtes étiquetées 1 a pour loi $Bin(L, \frac{1}{2})$, où L est la longueur du cycle, et est donc impair est probabilité $\frac{1}{2}$.
Ainsi, s'il existe, asymptotiquement, "beaucoup" de cycles qui imposent des contraintes indépendantes, on aura $P_{SAT}(\alpha) = 0$. Dans la preuve qui suit, on met en oeuvre cette idée en s'appuyant sur l'existence d'une composante connexe géante.

Proposition 1.2. Pour tout $\alpha > \frac{1}{2}$, $P_{SAT}(\alpha) = 0$.

Preuve. Considérons le graphe associé au système à N variables. On partage le graphe en deux parties, G_1, contenant $N - \lfloor\sqrt{N}\rfloor$ sommets, et G_2 contenant donc $\lfloor\sqrt{N}\rfloor$ sommets. (On pourrait prendre, au lieu $\lfloor\sqrt{N}\rfloor$, n'importe quelle suite (a_N) d'entiers vérifiant simplement $a_N = o(N)$ et $a_N \to \infty$). Puisque $N - \lfloor\sqrt{N}\rfloor \sim N$, alors pour N assez grand, le grand sous-graphe G_1 se situe lui-même au-dessus de la transition de percolation. On peut donc lui appliquer le théorème 1.1, qui nous fournit un réel $\delta > 0$ tel que, a.p.s., G_1 ait une composante connexe de cardinal au moins égal à δN.

Soit C cette composante connexe, et $|C|$ son cardinal. On considère x dans G_2, et y dans C. Le nombre de sommets de C auxquels x est relié a pour loi $Bin(|C|, p)$, où $p = 1 - (1 - \frac{\alpha}{N})^2$ (puisque chaque arête peut être étiquetée par 0 ou 1). On note que si x est relié à deux sommets y_1 et y_2 de C, alors il existe un cycle passant par x, y_1 et y_2. On souhaite donc minorer, asymptotiquement, $\mathbb{P}(n \geq 2)$, où n est le nombre de sommets de C auquel x est relié. On a $\mathbb{P}(n = 2) = \frac{|C|(|C|-1)}{2} p^2 (1-p)^{|C|-2}$ donc, comme asymptotiquement, $p \sim \frac{2\alpha}{N}$ et $\delta N \leq |C|$, on a, pour N grand : $\mathbb{P}(n = 2) \geq 2\alpha^2 \exp(-2\delta\alpha)$, donc

$\mathbb{P}(n \geq 2) \geq 2\alpha^2 \exp(-2\delta\alpha)$
Soit n', le nombre de points de G_2 reliés à au moins deux sommets de C. n' a pour loi $Bin(\lfloor\sqrt{N}\rfloor, \mathbb{P}(n \geq 2))$. La minoration de $\mathbb{P}(n \geq 2)$ montre que :

$$\forall d \in \mathbb{N}^*, \mathbb{P}(n' \leq d) \to 0 \tag{5}$$

Ainsi, a.p.s, il existe d cycles qui imposent chacun une contrainte qui est satisfiable avec probabilité $\frac{1}{2}$. De plus, ces contraintes sont indépendantes : en effet, on remarque que chacun de ces cycles contient une arête présente dans aucun des autres (c'est le cas des arêtes reliant un sommet x de G_2 à un sommet de G_1). Donc connaître les étiquettes des arêtes de certains de ces cycles ne donne pas d'information sur la parité du nombre d'étiquettes 1 présentes dans un des autres cycles. On a alors $P_{SAT} \leq \frac{1}{2^d}$, et c'est vrai pour tout d, ce qui achève la preuve. □

Remarque. On verra plus loin que $P_{SAT}(\alpha) \to 0$ lorsque $\alpha \to \frac{1}{2}^-$. Or P_{SAT} est décroissante en fonction de α : lorsque α augmente, chaque équation est présente avec plus grande probabilité ; si $\alpha < \alpha'$, on peut construire un espace de probabilité sur lequel deux systèmes aléatoires sont définis, l'un correspondant à notre modèle avec paramètre α, l'autre avec paramètre α', tel que le premier soit toujours sous-système de l'autre. La proposition 1.2 sera alors directement impliquée par le calcul de P_{SAT} pour $\alpha < \frac{1}{2}$. Néanmoins, on voit ici qu'on peut l'obtenir sans faire appel au calcul de P_{SAT} dans la zone $\alpha < \frac{1}{2}$ à l'aide d'arguments simples de percolation ; de plus, la preuve met en évidence que c'est la présence de cycles qui gêne la satisfiabilité.

1.3.2 Principe du calcul de P_{SAT} lorsque $\alpha < \frac{1}{2}$

Pour $\alpha < \frac{1}{2}$, il s'agit d'obtenir une expression explicite de P_{SAT}. Le résultat suivant montre, encore grâce au théorème 1.1, que la valeur de P_{SAT} est déterminée par les composantes connexes du graphe qui contiennent exactement un cycle :

Lemme 1.3. Supposons $\alpha < \frac{1}{2}$. Soit x un sommet du graphe, on note C_x la composante connexe qui le contient, et $|C_x|$ son cardinal. Alors la probabilité pour que C_x soit pluricyclique est $O((\log N)^4/N^2)$.
En particulier, a.p.s, toutes les composantes connexes du graphe contiennent au plus un cycle.

Preuve. L'algorithme du paragraphe 1.2 permet de générer un arbre dont les sommets sont les points de C_x. C_x est pluricyclique s'il existe encore deux arêtes supplémentaires au moins. Ceci est réalisé avec probabilité plus petite que $\binom{|C_x|}{2}^2 (\frac{\lambda}{N})^2$. Par le théorème 1.1, a.p.s, $|C_x| \leq a \log N$ où $a \in \mathbb{R}^{+*}$, d'où le résultat. □

On peut donc ignorer les éventuelles composantes connexes pluricycliques, qui disparaissent dans la limite thermodynamique. Une composante connexe ayant exactement un cycle est satisfiable avec probabilité $\frac{1}{2}$, et une composante connexe sans cycle est toujours satisfiable, on a donc :

$$P_{SAT}(\alpha) = \lim_{N\to\infty} \mathbb{E}(1/2^{n_c(N,\alpha)}) \tag{6}$$

où $n_c(N,\alpha)$ est le nombre de composantes contenant exactement un cycle. Il faut donc étudier la convergence en loi de $n_c(N,\alpha)$.

1.3.3 Estimation du nombre de cycles pour $\alpha < \frac{1}{2}$

Pour tout $L \geq 2$, on introduit $n_L(N,\alpha)$, le nombre de cycles de longueur L. On va montrer le résultat suivant :

Théorème 1.4. Lorsque $N \to \infty$, pour tout L, $(n_2, n_3, ..., n_L)$ converge en loi vers une suite de variables de Poisson indépendantes, de paramètres $\mu_2, ..., \mu_L$, avec $\mu_k = \frac{(2\alpha)^k}{2k}$.

Si X est une variable aléatoire à valeurs dans $\mathbb{N}$, on définit le moment factoriel d'ordre p de X par $\mathbb{E}([X]_p) = \mathbb{E}(X(X-1)...(X-p+1))$. Plus généralement, pour k variables aléatoires X_1, ..., X_k, on peut définir leurs moments factoriels conjugués $\mathbb{E}([X_1]_{p_1}...[X_k]_{p_k})$. Si les X_i sont k variables de Poisson indépendantes de paramètres $\mu_1, ..., \mu_k$, alors ces moments valent $\prod_{i=1}^{k} \mu_i^{p_i}$. On va prouver que les moments factoriels conjugués des n_L convergent vers ceux des variables voulues, puis qu'une telle convergence implique la convergence en loi.

Lemme 1.5. Pour tout $k \in \mathbb{N}^*$, pour tous $2 \leq L_1 < ... < L_k$ dans $\mathbb{N}^*$, et $p_1, ..., p_k$ dans $\mathbb{N}$, on a :

$$\mathbb{E}([n_{L_1}]_{p_1}, ...[n_{L_k}]_{p_k}) \to_{N\to\infty} \prod_{i=1}^{k} (\frac{(2\alpha)^{L_i}}{2L_i})^{p_i}.$$

Preuve. (i) Fixons $L \geq 2$. En raisonnant par récurrence sur p, on voit qu'on a, pour tout $p \in \mathbb{N}$:

$$[n_L]_p = n_L(n_L-1)...(n_L-p+1) = \sum_{*} 1_{C_1}...1_{C_p} + O(\sum_{**} 1_{C_1}...1_{C_m})$$

où 1_H est l'indicatrice de la présence d'un sous-graphe H ; la sommation $(*)$ se fait sur les p-uplets de cycles de longueur L n'ayant aucune arête commune, et la sommation $(**)$ se fait sur les m-uplets de cycles de longueur L, $m = 1, ..., p$, les m cycles étant deux à deux distincts mais tels qu'au moins d'entre eux ont une arête commune.
En effet, supposons ce résultat vrai au rang p, et multiplions-le par $n_L = \sum 1_C$ (la somme portant sur tous les cycles de longueur L). Multiplier par 1_C donne trois types de termes :

- $1_C \times 1_{C_1}...1_{C_p}$ où $(C, C_1, ..., C_p)$ sont deux à deux disjoints. La somme de ces termes vaut $\sum_* 1_{C_1}...1_{C_{p+1}}$
- $1_C \times 1_{C_1}...1_{C_p}$ où $(C_1, ..., C_p)$ sont deux à deux disjoints et C est l'un des C_i. La somme de ces termes vaut $p\sum_* 1_{C_1}...1_{C_p}$.
- Dans les autres cas, au moins deux des cycles $C, C_1, ..., C_p$ ou $C, C_1, ..., C_m$ ont une arête commune sans être égaux. On obtient des termes $1_{C_1}, ..., 1_{C_{m'}}$ analogues à ceux de $(**)$, et on peut borner le nombre de fois où chacun apparaît par une constante dépendant uniquement de p+1.

En retranchant $p[n_L]_p$ aux deux membres, on obtient donc le résultat au rang p+1.

(ii) On en déduit :

$$\mathbb{E}([n_{L_1}]_{p_1}...[n_{L_k}]_{p_k}) = \sum_{*'} \mathbb{E}((1_{C_{1,1}}...1_{C_{1,p_1}})...(1_{C_{k,1}}...1_{C_{k,p_k}})) + O(\sum_{**'} \mathbb{E}(1_{C_1}...1_{C_m})) \quad (7)$$

où la sommation $(*')$ se fait sur les $(p_1 + ... + p_k)$-uplets de cycles n'ayant aucune arête commune dont exactement p_i ont une longueur L_i pour tout i, et la sommation $(**')$ se fait sur les m-uplets de cycles de longueur dans $\{L_1, ..., L_k\}$, $m = 1, ..., p_1 + ... + p_k$, les m cycles étant deux à deux distincts mais tels qu'au moins d'entre eux ont une arête commune.

(iii) En convenant que $L_0 = 0$, la première somme dans (7) vaut :

$$\sum_{*'} = \prod_{i=1}^{k} \frac{(N - ((L_0 + ... + L_{i-1})p)...(N - (L_0 + ... + L_i)p + 1)}{(2L_i)^{p_i}} (\frac{2\alpha}{N})^{L_i p_i}$$

ce qui montre que

$$\sum_{*'} \to_{N\to\infty} \prod_{i=1}^{k} (\frac{(2\alpha)^{L_i}}{2L_i})^{p_i}$$

Dans chaque terme $\mathbb{E}(1_{C_1}...1_{C_m})$ de la somme $(**')$, on obtient des espérances d'indicatrices de réunions de cycles, contenant au moins deux cycles s'intersectant. Si on regroupe ensuite les termes qui correspondent à des graphes identiques à renumérotation près des sommets, les "classes" de graphes ainsi obtenues sont en nombre borné indépendamment de N. De plus, le nombre moyen d'éléments d'une classe C donnée présents dans le graphe est

$$k(C) = \binom{N}{S} (\frac{2\alpha}{N})^A \frac{S!}{Aut(C)}$$

où A et S sont respectivement les nombres d'arêtes et de sommets de tout élément de la classe C, et Aut(C) le nombre d'automorphismes de tout élément de C (un automorphisme de graphe étant une bijection des sommets respectant la structure de graphe i.e. transformant toute arête en arête). Ces éléments étant union de cycles et de sous-graphes connexes

pluricycliques, on a $A - S \geq 1$, ce qui montre $k(A,S)_{N\to\infty} \to 0$. $\square$

Lemme 1.6. Si des variables $X_1, ..., X_k$ à valeurs dans $\mathbb{N}$ vérifient $\mathbb{E}([X_1]_{p_1}...[X_k]_{p_k}) \to \prod_{i=1}^{k} \lambda_i^{p_i}$ pour tous $p_1, ..., p_k$ dans $\mathbb{N}$, alors $(X_1, ..., X_k)$ converge en loi vers une suite $(Z_1, ..., Z_k)$ de variables de Poisson indépendantes dont les paramètres sont $\lambda_1, ..., \lambda_k$.

Preuve. (i) Supposons d'abord k = 1. Par le théorème de Banach-Alaoglu, on peut extraire de toute sous-suite de X_1 une suite qui converge en loi. Il suffit donc de montrer que toutes ces sous-suites convergent vers la même limite.

On considère donc Z_1, limite en loi d'une suite extraite $(X_{1,\phi(N)})$ de X_1. On se donne $p \in \mathbb{N}$. Puisque les moments factoriels de $X_{1,\phi(N)}$ convergent, les moments $\mathbb{E}(X_{1,\phi(N)})^p$ convergent aussi. En particulier, $C_p := \sup_N \mathbb{E}(X^p_{1,\phi(N)})$ est fini. On a donc, pour tout $i \in \mathbb{N}^*$, $\mathbb{P}(X_{1,\phi(N)} = i)i^p \leq \frac{C_{p+2}}{i^2}$; par convergence dominée, il s'ensuit $\mathbb{E}(X^p_{1,\phi(N)}) \to \mathbb{E}(Z_1^p)$. Z_1 a donc des moments à tout ordre, et ils sont égaux à ceux d'une variable de Poisson de paramètre λ_1. La fonction caractéristique d'une loi de Poisson pouvant se développer en série entière, Z_1 suit donc une loi de Poisson de paramètre λ_1.

(ii) On suppose le résultat vrai jusqu'au rang $k-1$ et on le montre au rang k. On distingue deux cas :
1) $\lambda_k = 0$. En prenant $p_1 = ... = p_{k-1} = 0$ et $p_k = 1$ dans l'hypothèse de l'énoncé, on obtient alors $\mathbb{E}(X_{k,N}) \to 0$, ce qui implique que pour tout m, $P(X_{k,n} = m) \to \delta_{m,0}$. L'hypothèse de récurrence permet alors de conclure directement.
2) $\lambda_k > 0$. On se donne $r_1, ...r_k$ dans $\mathbb{N}$. Notant $A_N = \{X_{1,N} = r_1, ..., X_{k-1,N} = r_{k-1}\}$, on a $P(X_{1,N} = r_1, ..., X_{k,N} = r_k) = P(X_{k,N} = r_k|A_N)P(A_N)$. Il s'agit donc de montrer que $P(X_{k,N} = r_k|A_N) \to e^{\lambda_k}\frac{\lambda_k^{r_k}}{r_k!}$.
Soit $p_k \in \mathbb{N}$. On effectue un changement de probabilité en posant, pour tout événement A, $P'(A) = \frac{\mathbb{E}([X_{k,N}]p_k|A)P(A)}{\mathbb{E}([X_{k,N}]p_k)}$. Dans ce nouvel espace, l'espérance d'une variable est donnée par $\mathbb{E}'(X) = \frac{\mathbb{E}(X[X_{k,N}]_{p_k})}{\mathbb{E}([Xk,N]_{p_k})}$. Ainsi, la convergence des moments factoriels vers les $\prod \lambda_i^{r_i}$ au rang k-1 est vérifiée également pour P'. Par hypothèse de récurrence, on a donc $P'(A_N) \to \prod_{i=1}^{k-1} e^{\lambda_i}\frac{\lambda_i^{r_i}}{r_i!}$.
$P(A_N)$ et $P'(A_N)$ tendent donc vers la même limite non nulle. Revenant à la définition de P', ceci implique que $\mathbb{E}([X_{k,N}]_{p_k}|A_N) - \mathbb{E}([X_{k,N}]_{p_k}) \to 0$, i.e. $\mathbb{E}([X_{k,N}]_{p_k}|A_N) \to \lambda_k^{p_k}$
Ceci étant vrai pour tout p_k, on peut appliquer le résultat au rang 1 à $X_{k,N}$ conditionné par rapport à A_N ; il vient $P(X_{k,N} = r_k|A_N) \to e^{\lambda_k}\frac{\lambda_k^{r_k}}{r_k!}$, ce qui termine la preuve. $\square$

1.3.4 Conclusion

L'égalité (6) et les résultats précédents montrent que $P_{SAT} = \mathbb{E}(2^{-\sum_{l\geq 2} Z_l})$, où les Z_l sont des variables de Poisson indépendantes de paramètres $\frac{(2\alpha)^L}{2L}$. Par le théorème de convergence dominée, et par l'indépendance des Z_l, on a donc : $P_{SAT} = \prod_2^\infty \mathbb{E}(2^{-Z_l})$. On obtient alors $P_{SAT}(\alpha) = e^{\frac{\alpha}{2}}(1-2\alpha)^{\frac{1}{4}}$. Enfin, comme P_{SAT} décroît en fonction de α, on a $P_{SAT}(\frac{1}{2}) = 0$.

Théorème 1.7. La probabilité de satisfiabilité lorsque $N \to \infty$ est donnée par :

$$P_{SAT}(\alpha) = \begin{cases} e^{\frac{\alpha}{2}}(1-2\alpha)^{\frac{1}{4}} & \text{si } \alpha < \frac{1}{2}, \\ 0 & \text{si } \alpha \geq \frac{1}{2} \end{cases}$$

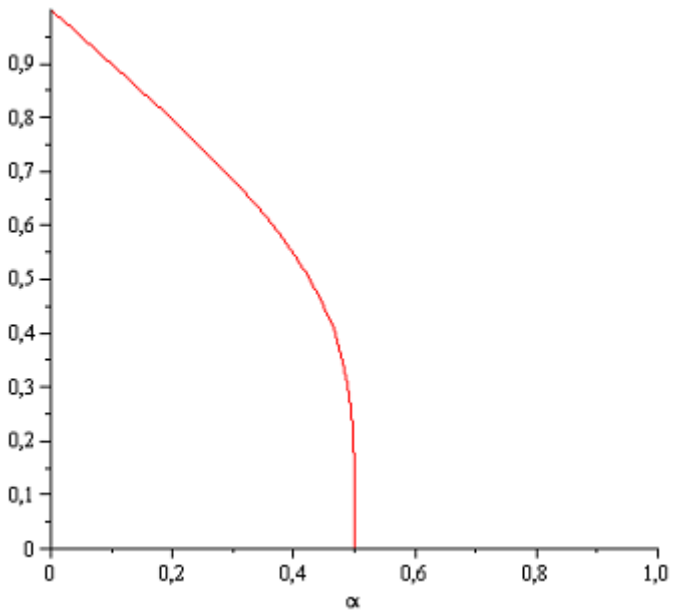

FIGURE 2 – La fonction $\alpha \mapsto P_{SAT}(\alpha)$ pour 2-XORSAT

En particulier, P_{SAT} est non nulle pour $\alpha < \frac{1}{2}$ et nulle pour $\alpha \geq \frac{1}{2}$. Il y a de plus une tangente verticale à gauche en $\alpha = \frac{1}{2}$. Cependant, P_{SAT} dépend continument de α.

2 p-XORSAT, $p \geq 3$

2.1 Comportement général

Nous étudions à présent la situation $p \geq 3$. De manière analogue au cas précédent, on associe au système un hypergraphe, où les hyperarêtes relient p sommets.

Le comportement du système en fonction de α est assez différent de celui de 2-XORSAT. D'une part, la transition de satisfiabilité ne correspond plus à une transition de percolation. D'autre part, il existe en fait deux transitions : la transition de satisfiabilité, et une deuxième, plus subtile, qui concerne la structure de l'ensemble des solutions.
Nous introduirons dans cette partie un algorithme qui nous permettra de calculer P_{SAT} tout en comprenant pourquoi la percolation ne joue plus de rôle dans la transition. Enfin nous définirons de nouveaux paramètres d'ordre qui nous permettront d'expliquer la nature de la deuxième transition.

2.2 Le core de l'hypergraphe

Avant toute chose, décrivons un peu ce que devient le modèle pour des systèmes d'équations à p variables. Chaque équation est toujours présente avec une certaine probabilité q, et ces évènement sont tous indépendants. On voudrait fixer q de sorte que le nombre moyen d'équations soit $M \simeq \alpha N$ pour un paramètre α qu'on gardera fixe lorsqu'on étudiera ensuite la limite $N \to \infty$. On doit avoir alors $2q\binom{N}{p} \simeq \alpha N$, on choisit donc $q = \frac{\alpha p!}{2N^{p-1}}$.
On remarque alors que pour p>2, la probabilité que deux équations avec les mêmes p variables et des seconds membres différents soient présentes dans le système est inférieure à $\binom{N}{p}(\frac{\alpha p!}{2N^{p-1}})^2 \sim \frac{\alpha p!}{4N^{p-2}} = o(1)$. On peut donc, par souci de simplicité, modifier un peu le modèle (sans changer P_{SAT}) : une équation contenant p variables données sera présente avec probabilité $\frac{\alpha p!}{N^{p-1}}$, et son second membre sera 0 ou 1 avec probabilité $\frac{1}{2}$, les seconds membres des équations étant indépendants des membres de gauche.

2.2.1 Algorithme d'effeuillage

On avait vu dans l'étude de 2-XORSAT, que l'étude de la satisfiabilité du système global se ramenait à l'étude des cycles du graphes sous-jacents. On aimerait procéder de même, et ramener l'étude de la satisfiabilité dans p-XORSAT à l'étude d'un sous-hypergraphe plus simple, ou du moins plus petit. L'algorithme d'effeuillage permet de répondre à cette question. Supposons que dans le système, une des variables, notons-la x_i, n'apparaisse que dans une seule équation (E). Alors le système global est satisfiable si et seulement si le système dans lequel on a supprimé l'équation (E) et la variable x_i est satisfiable : en effet, pour toute solution de ce sous-système, il existe un choix de la valeur de x_i satisfiant l'équation (E) (il suffit de reporter la valeur des autres variables dans l'équation (E)).
On peut itérer ce procédé tant qu'il existe une variable qui n'apparaît que dans une seule

équation. La transformation sous-jacente sur l'hypergraphe G associé au système consiste, tant que cela est possible, à choisir un sommet de degré 1 au hasard (uniformément parmi ces sommets), et à supprimer l'unique hyperarête le contenant. Cet algorithme s'arrête lorsque tous les sommets de l'hypergraphe restant sont de degré ≥ 2. L'hypergraphe obtenu est en fait indépendant des choix faits lors de l'algorithme d'effeuillage : il s'agit en effet du plus grand grand (au sens de l'inclusion) sous-hypergraphe de G dans lequel tous les sommets sont de degrés ≥ 2 : si H est un sous-hypergraphe de G dont tous les sommets sont de degré ≥ 2 alors à aucun moment l'algorithme d'effeuillage ne pourra supprimer d'arête de H (il est aisé de le voir par récurrence sur le nombre d'étapes de l'algorithme). Donc, le graphe obtenu en fin d'algorithme contient tous les sous-hypergraphes de G dont tous les sommets sont de degré ≥ 2. De plus, puisque l'algorithme est arrêté, tous les sommets de l'hypergraphe obtenu sont de degré ≥ 2, donc c'est bien le plus grand sous-hypergraphe de G dont tous les sommets de degré ≥ 2, ce qui montre qu'il est indépendant des choix dans l'algorithme. L'hypergraphe obtenu est alors nommé core de G (ou 2-core, parce qu'il s'agit du plus grand sous-hypergraphe de G dont les sommets sont de degré ≥ 2). Le système initial est alors satisfiable si et seulement si le système associé au core est satisfiable.

Dans les deux parties qui suivent, nous chercherons à étudier le structure du core pour l'hypergraphe aléatoire étudié : nombre de sommets, d'arêtes, degrés des sommets...L'idée est qu'on peut calculer en espérance la variation du nombre N_d de sommets de degré d en une étape de l'algorithme en fonction de $N_0, N_1..., N_d, ...$ On pourra alors montrer que les N_d évoluent de manière proche de la solution d'un système d'équations différentielles. La partie 2.2.2 présente un théorème général d'approximation d'une suite de variables aléatoires grâce à des équations différentielles, et la partie 2.2.3 applique ces résultats à l'étude du core.

2.2.2 Evolution de variables aléatoires et équation différentielle

Dans cette section, nous présentons donc une approche générale permettant d'avoir une estimation asymptotique d'une suite de variables aléatoires par une équation différentielle. Le théorème présenté ainsi que sa démonstration proviennent largement de [4], l'article contenant également de multiples applications de ce résultat, notamment pour l'étude des graphes.

Soient S un ensemble, $Q = (Q_0, Q_1, ..., Q_t...)$ un processus aléatoire à valeurs dans S et H_t la tribu engendrée par $(Q_0, Q_1, ..., Q_t)$. Soient $y_l : S \longrightarrow \mathbb{R}$, pour $1 \leq l \leq a$, des fonctions. On note $Y_l(t)$ la variable aléatoire $y_l(Q_t)$.
Nous considérons une suite de tels procédés $Q^{(N)}$ indexés par N. Notons que $S = S^{(N)}$ dépend aussi de N. Dans la suite, la dépendance en N sera implicite.
Enfin pour un ouvert borné connexe $D \subset \mathbb{R}^{a+1}$, on introduit un temps d'arrêt

$T_D = \min\{t/(\frac{t}{N}, \frac{Y_1(t)}{N}, ..., \frac{Y_a(t)}{N}) \notin D\}$ et on dit qu'une fonction f est L-lipschitzienne sur D si

$$\forall u, v \in D \, |f(u) - f(v)| \leq L * \max_{1 \leq i \leq a+1} (|u_i - v_i|)$$

Nous pouvons alors énoncer le théorème :

Théorème 2.1. Avec les notations précédentes, on fait les hypothèses

(i) (Variations bornées) $\exists \beta > 0 \ / \forall t < T_D, \ \max_{1 \leq l \leq a} (Y_l(t+1) - Y_l(t)) \leq \beta$

(ii) (Equation de transition) Pour une fonction $\lambda(N) = o(1)$ on a $|E(Y_l(t+1) - Y_l(t)|H_t) - f_l(\frac{t}{N}, \frac{Y_1(t)}{N}, ...\frac{Y_a(t)}{N})| \leq \lambda$ où les f_l sont des fonctions L-lispchitziennes sur D pour un certain L.

Alors on a :

(a) Pour tout $(0, z_1, .., z_a) \in D$, le système d'équations différentielles $Z_l'(t) = f_l(t, Z_1(t), ...Z_a(t))$ $Z_l(0) = z_l \ 1 \leq l \leq a$ admet une solution Z qui peut être définie tant que Z n'a pas atteint la frontière de D.

(b) Pour une constante C assez grande, avec probabilité $1 - O(a\frac{\beta}{\lambda}\exp(-N\frac{\beta^3}{\lambda^3}))$ on a $Y_l(t) = NZ_l(t/N) + O(\lambda N)$ avec Z solution du système différentiel avec $Z_l(0) = \frac{Y_l(0)}{N}$ le O étant uniforme pour $0 \leq t \leq \sigma N$ où $\sigma = \inf\{t/d((t, Z_1(t), ..., Z_a(t)), Fr(D)) \leq C\lambda\}$

Preuve. Tout d'abord, l'existence des solutions est assurée du fait de l'hypothèse de lipschitziennité des fonctions par le théorème de Cauchy-Lipschitz.
Pour simplifier les notations, supposons d'abord que $a = 1$. On introduit $w = \lceil \frac{N\lambda}{\beta} \rceil$. La démonstration se fait en deux temps : on montre que $Y(t+w) - Y(t)$ est concentré autour de $wf(t, \frac{Y(t)}{N})$ puis on estime l'erreur entre Y et Z aux points kw, $0 \leq k \leq i_0, \ i_0 = \lceil \frac{\sigma N}{w} \rceil$

- Etape 1 : Nous présentons d'abord un lemme qui permettra de contrôler $Y(t+w) - Y(t)$ à l'aide de surmartingales :

Lemme 2.2. Soit X_i une surmartingale par rapport à H_t avec $|X_i - X_{i-1}| \leq c$.
Alors $P(X_t \geq \alpha) \leq \exp(-2\frac{\alpha^2}{tc^2})$

Preuve. Prenons h>0. Nous optimiserons la valeur de h plus tard. On utilise l'inégalité de Markov :

$$P(X_t \geq \alpha) = P(e^{hX_t} \geq e^{h\alpha}) \leq e^{-h\alpha} E(e^{hX_t}) \tag{8}$$

Or $E(e^{hX_t}) = E(e^{hX_{t-1}} e^{h(X_t - X_{t-1})}) = E(e^{hX_{t-1}} E(e^{h(X_t - X_{t-1})}|H_{t-1}))$ puisque $e^{X_{t-1}}$ est H_{t-1}-mesurable et intégrable (par hypothèse de récurrence). A présent par convexité de

$x \to e^{hx}$ et comme $|X_t - X_{t-1}| \leq c$ on a

$$E(e^{h(X_t - X_{t-1})}|H_{t-1})) \leq E(\frac{e^{hc} + e^{-hc}}{2} + \frac{(X_t - X_{t-1})}{2c}|H_{t-1})) \leq \cosh(hc) \leq e^{\frac{(hc)^2}{2}} \quad (9)$$

car X_t est une surmartingale. On a donc $E(e^{hX_t}) \leq e^{\frac{(hc)^2}{2}} E(e^{hX_{t-1}})$ puis par récurrence $E(e^{hX_t}) \leq e^{t\frac{(hc)^2}{2}}$, ce qui donne $P(X_t \geq \alpha) \leq e^{t\frac{(hc)^2}{2}} e^{h\alpha}$. La valeur optimale pour h est $h = \frac{\alpha}{tc^2}$ ce qui donne l'inégalité du lemme.

A présent, estimons $Y(t+w) - Y(t)$ à l'aide du lemme :

$$E(Y(t+k+1) - Y(t+k)|H_{t+k}) = f(\frac{t+k}{N}, \frac{Y(t+k)}{N}) + O(\lambda) = f(\frac{t}{N}, \frac{Y(t)}{N}) + O(\lambda + \frac{k\beta}{N}) \quad (10)$$

ceci grace au caractère lipschitzien de f et au fait que les variations de Y sont bornées par β. Donc pour K assez grand, si $g(N) = K(\lambda + \frac{w\beta}{N})$ alors $Y(t+k) - Y(t) - kf(\frac{t}{N}, \frac{Y(t)}{N}) - kg(N)$ est une surmartingale en k par rapport à $(H_{t+k})_{k=0\ldots w}$. Les variations de cette surmartingale sont bornées par une constante κ, et on en déduit par le lemme :

$$P(Y(t+w) - Y(t) - wf(\frac{t}{N}, \frac{Y(t)}{N}) - wg(N) \geq \kappa\sqrt{2w\alpha} \leq e^{-\alpha} \quad (11)$$

L'inégalité étant valable pour tout $\alpha > 0$, on choisit par la suite $\alpha = N\frac{\lambda^3}{\beta^3}$. De plus, si K est assez grand en remplaçant g(n) par -g(n) on obtient une sous-martingale et par le lemme une minoration analogue de $Y(t+w) - Y(t) - wf(\frac{t}{N}, \frac{Y(t)}{N})$:

$$P(|Y(t+w) - Y(t) - wf(\frac{t}{N}, \frac{Y(t)}{N})| \geq wg(N) + \kappa\sqrt{2w\alpha}) \leq 2e^{-\alpha} \quad (12)$$

- Etape 2 : Nous allons contrôler l'écart entre Y et Z aux points $kw,\ 0 \leq k \leq i_0,\ i_0 = \lceil \frac{\sigma N}{w} \rceil$. Comme les variations de Y et de Z sont en $O(\lambda N)$ sur une distance de w cela suffira pour prouver le théorème. On va procéder par récurrence : on a

$$|Y((k+1)w) - NZ(\frac{(k+1)w}{N})| \leq |Y(kw) - NZ(\frac{kw}{N})| + |Y((k+1)w) - Y(kw) - wf(\frac{kw}{N}, \frac{Y(kw)}{N})| +$$

$$|NZ(\frac{(k+1)w}{N}) - NZ(\frac{kw}{N}) - wf(\frac{kw}{N}, Z(\frac{kw}{N})| + |wf(\frac{kw}{N}, \frac{Y(kw)}{N}) - wf(\frac{kw}{N}, Z(\frac{kw}{N}))| \quad (13)$$

L'étape précédente permet d'affirmer qu'avec probabilité $1 - O(\exp(-\alpha))$ le deuxième terme est $< Bw\lambda$ pour une constante universelle B.
Par l'égalité des accroissements finis, il existe $c \in]kw, (k+1)w[$ tel que $NZ(\frac{(k+1)w}{N}) - NZ(\frac{kw}{N}) = wZ'(c)$ et comme f est L lipschitzienne le troisième terme est $\leq L\frac{w^2}{N}$

Finalement on a avec probabilité $1 - O(\exp(-\alpha))$:

$$|Y((k+1)w) - NZ(\frac{(k+1)w}{N})| \leq |Y(kw) - NZ(\frac{kw}{N})|(1 + \frac{wL}{N}) + Bw\lambda + L\frac{w^2}{N} \quad (14)$$

Il s'ensuit, par récurrence et par le fait que $Z(0) = \frac{Y(0)}{N}$, que avec probabilité $1 - O(k\exp(-\alpha)$

$$|Y(kw) - NZ(\frac{kw}{N})| \leq (\frac{NB\lambda}{L} + w)((1 + \frac{Lw}{N})^k - 1) = B_k \quad (15)$$

Comme $k \leq \frac{\sigma N}{w}$, $(1 + \frac{Lw}{N})^k$ est un O(1) uniformément en k, on a majoré l'erreur par $B_k = O(\lambda N)$, uniformément en k. Le résultat se prouve de la même manière si le nombre d'équations est $a > 1$ (avec a pouvant dépendre de n), il suffit de formuler comme hypothèse de récurrence $\mathbb{P}(|Y_l(kw) - NZ_l(\frac{kw}{N})| \geq B_k) = O(ak\exp(-\alpha))\ \forall 1 \leq l \leq a$. □

2.2.3 Application aux propriétés du core

Nous allons à présent chercher à appliquer le théorème précédent pour calculer l'évolution du graphe au cours de l'algorithme d'effeuillage, plus précisemment l'évolution de la distribution des degrés dans le graphe. Le théorème 2.1 va nous permettre de montrer que celle-ci se concentre autour de la solution d'une équation différentielle, jusqu'au temps T où le nombre de sommets de degré 1 atteint 0. Le graphe obtenu au temps T étant le core, nous en déduirons les caractéristiques du core du graphe : nombre de sommets et d'arêtes, distribution des degrés... en prenant tout simplement la valeur finale (i.e. au temps T) dans la solution de l'équation différentielle.

Notons N_d le nombre de sommets de degré d dans l'hypergraphe aléatoire (à N sommets) et M son nombre d'arêtes. Toutes sont des variables aléatoires dépendant de t le nombre d'étapes de l'algorithme d'effeuillage depuis l'hypergraphe initial. On cherche à appliquer le théorème 2.1 à $Z(t) = (N_0(t), N_1(t), ..., M(t))$. Pour pouvoir exploiter le théorème 2.1. nous avons besoin de deux choses : il faut vérifier que les conditions intiales sont concentrées autour de leur valeur moyenne, ce qui permettra d'aboutir à une concentration autour d'une solution d'équation différentielle avec conditions initiales déterministes (là où dans le théorème 2.1 la condition initiale est une variable aléatoire), et d'autre part vérifier les hypothèses du théorème 2.1 Parmi celles-ci, la condition de variations bornées est trivialement satisfaite : au cours d'une étape de l'algorithme de transition, on retire une hyperarête, donc les variations des N_d sont bornées par p (et celle de M par 1). L'étude se réduit à trouver une équation de transition approchée, et à majorer en espérance l'erreur commise.

Nous commençons par étudier le premier point, le plus simple : la concentration des conditions intiales.

Proposition 2.3. Pour tout d, $\mathbb{E}(N_d) = Ne^{-\alpha p}\frac{(\alpha p)^d}{d!} + O(1)$ et $Var(N_d) = O(N)$. De plus $\mathbb{E}(M) = \alpha N + O(1)$ et $Var(M) = O(N)$

Preuve. Tout d'abord la loi de M est $Bin(\binom{N}{p}, \frac{\alpha p!}{N^{p-1}})$ donc sa moyenne est $\binom{N}{p}\frac{\alpha p!}{N^{p-1}} = \alpha N + O(1)$ et sa variance est $\binom{N}{p}\frac{\alpha p!}{N^{p-1}}(1 - \frac{\alpha p!}{N^{p-1}}) = O(N)$.
Soit maintenant x un sommet du graphe. Calculons la probabilité qu'il soit de degré d : son degré a pour loi $Bin(\binom{N-1}{p-1}, \frac{\alpha p!}{N^{p-1}})$ donc la probabilité pour qu'il soit de degré d est

$$\binom{\binom{N-1}{p-1}}{d}(\frac{\alpha p!}{N^{p-1}}))^d(1-\frac{\alpha p!}{N^{p-1}}))^{\binom{N-1}{p-1}-d} = (\frac{(\frac{N^{p-1}}{(p-1)!})^d}{d!} + O(N^{(p-1)d-1}))(\frac{\alpha p!}{N^{p-1}}))^d(e^{-\alpha p} + O(\frac{1}{N}))$$

$$= e^{-\alpha p}\frac{(\alpha p)^d}{d!} + O(\frac{1}{N})$$

En sommant sur tous les sommets du graphe, on obtient le résultat annoncé.
On termine par le calcul de la variance de N_d. Soit un couple de sommets x et y. La probabilité qu'il y ait une hyperarête entre x et y est majorée par $\binom{N-2}{p-2}\frac{\alpha p!}{N^{p-1}}$ c'est donc un $O(\frac{1}{N})$. Donc, la probabilité pour qu'ils soient tous deux de degré d est, à un $O(\frac{1}{N})$ près, la probabilité pourqu'il soient tous deux dans exactement d hyperarêtes parmi celles qui ne contiennent pas à la fois x et y. Cette probabilité est $(\binom{\binom{N-2}{p-1}}{d})(\frac{\alpha p!}{N^{p-1}})^d(1 - \frac{\alpha p!}{N^{p-1}})^{\binom{\binom{N-2}{p-1}}{d}-d})^2 = e^{-2\alpha p}(\frac{(\alpha p)^d}{d!})^2 + O(\frac{1}{N})$
En sommant sur les couples (x,y) avec $x \neq y$, on obtient que le moment d'ordre deux est $N^2e^{-2p\alpha}(\frac{(p\alpha)^d}{d!})^2 + O(N)$ (dans un développement en indicatrices, les couples (x,x) n'ont qu'une contribution O(N) au second moment) donc la variance est un O(N). □

Avant de commencer l'étude de l'équation de transition, nous avons encore besoin d'un autrerésultat préliminaire sur le degré maximal dans le graphe.

Lemme 2.4. A.p.s. il n'y a aucun sommet de degré >log(N) lorsque l'hypergraphe aléatoire a N sommets.

Preuve. Soit d>log(N). La probabilité qu'un sommet soit de degré d est $\binom{\binom{N-1}{p-1}}{d}(\frac{\alpha p!}{N^{p-1}}))^d(1-\frac{\alpha p!}{N^{p-1}}))^{\binom{N-1}{p-1}-d} \leq \frac{\alpha^d}{d!} \leq C\frac{\alpha^{log(N)}}{(log(N))!} = o(N^{-k})$ pour tout k par Stirling. En sommant sur $\log(N) \leq d \leq \binom{N-1}{p-1}$ puis en sommant sur tous les sommets, on obtient que la probabilité qu'un sommet ait un degré » > >log(N) est un o(1). □

Nous pouvons à présent chercher à déterminer les équations de transition pour les variables $N_d(t)$. Ces calculs sont faits dans [5] avec une approche un peu différente de la

notre ([5] se base sur une approximation par un processus de branchement), avec toutefois la proposition 2.5. en commun (proposition 7.2 dans [5]) A la t-ième étape de l'algorithme, on choisit au hasard un sommet de degré 1 et on supprime son unique hyperarête. Les sommets de cette hyperarête voient tous leur degré diminuer de 1, les autres conservent leur degré. La variation $\Delta N_d(t) = N_d(t+1) - N_d(t)$ de sommets de degré d est donc simplement le nombre de sommets de degré d+1 dans l'hyperarête moins le nombre de sommets de degré d. Le problème est qu'a priori on ne connait pas bien la loi de l'hypergraphe G(t) après t étapes de l'algorithme. Toutefois, l'algorithme d'effeuillage possède la propriété suivante, fondamentale pour le calcul des équations de transitions :

Proposition 2.5. Conditionnellement à la distribution des degrés $\mathbf{d} = (N_0, N_1..)$ de ses sommets, l'hypergraphe G(t) après t étapes de l'algorithme d'effeuillage est uniformément distribué dans l'ensemble $G(N, \mathbf{d})$ des hypergraphes à N sommets et de distribution de degré $\mathbf{d}$.

Preuve. La preuve se fait par récurrence sur t. Le graphe initial G(0) est bien sûr uniformément distribué conditionnellement aux degrés, il est même uniformémemnt distribué, conditionnellement au nombre d'arêtes M(0), dans l'ensemble $G(N, M(0))$ des hypergraphes à N sommets et M(0) arêtes.
Supposons G(t) uniformément distribué conditionnellement aux degrés. Soit G hypergraphe de distribution de degrés $(N_0, N_1, ...)$, on cherche à connaître de combien d'hypergraphes G' de distribution de degrés $(N_0^{'}, N_1^{'}, ...)$ il peut provenir en une étape de l'algorithme d'effeuillage. Si $N_1^{'} = 0$ ce nombre est 0, puisqu'alors l'algorithme s'arrête au temps t. Sinon, pour reconstituer G' à partir de G il faut créer une nouvelle arête de sorte à obtenir la bonne distribution de degrés. L'arête doit donc contenir $N_0 - N_0^{'}$ sommets de degré 1 dans G, donc $N_1 - N_1^{'} + N_0 - N_0^{'}$ sommets de degré 2 dans G, $N_2 - N_2^{'} + N_1 - N_1^{'} + N_0 - N_0^{'}$ sommets de degré 2 etc...Si le nombre total de sommets que doit contenir la nouvelle hyperarête est différent de p le nombre de graphe G' est 0, sinon c'est $\prod_i \binom{N_{i+1}}{N_0 - N_0^{'} + ... + N_i - N_i^{'}}$ (on remarque que seuls un nombre fini de facteurs sont différents de 1). Chacun de ces hypergraphes a une probabilité $\frac{N_0 - N_0^{'}}{N_1^{'}}$ de donner le graphe G (l'hypergraphe G' ne peut donner G que si on supprime l'arête qu'on a construite pour passer de G à G' (puisque celle-ci n'était pas dans G) ; celle-ci contenant $N_0 - N_0^{'}$ sommets de degré 1, elle est choisie pour être supprimée avec la probabilité indiquée).
Conclusion de ces considérations : si G(t) est uniformément distribué, la probabilité, conditionnellement à ce que la distribution des degrés de G(t) soit $\mathbf{d}$, d'obtenir un certain hypergraphe G à l'instant t+1 ne dépend que de la distribution de degrés de G, ceci est de plus vrai pour tout $\mathbf{d}$. La probabilité $\mathbb{P}(G(t+1) = G)$ est la somme des $\mathbb{P}$(la distribution des degrés de G(t) est $\mathbf{d}$ et $G(t+1) = G$). Les membres de la somme ne dépendant que de la distribution des degrés de G, donc tous les hypergraphes ayant la même distribution

de degré ont la même probabilité à l'instant t+1, c'est-à-dire G(t+1) est uniformément distribué conditionnellement à sa distribution des degrés. □

On peut à présent chercher une équation de transition approchée pour la suite des degrés. Il s'agit d'estimer les variations ΔN_d de chaque nombre de sommets de degré d entre les instants t et t+1, conditionnellement à la distribution des degrés à l'instant t. On a vu que la variation de N_d est la différence du nombre de sommets de degré d+1 dans l'hyperarête supprimée à l'instant t et du nombre de sommets de degrés d dans cette même hyperarête. L'hyperarête a été choisie en choisissant aléaoirement un sommet x de degré 1 (avec probabilité uniforme) et en supprimant son arête, elle contient donc au moins un sommet de degré 1. Il y a p-1 autres sommets dans l'hyperarête, donc les degrés peuvent être quelquonques. La probabilité qu'un sommet de degré d appartienne à une hyperarête choisie aléatoirement est $p_d = \frac{d}{M}$ car chaque sommet de degré d appartient à d hyperarêtes. Si la loi du degré des voisins du sommet x de degré 1 dans l'hyperarête était la même que dans une hyperarête choisie au hasard, on aurait donc :

$$\mathbb{E}(\Delta N_d(t)|(N_0,...)) = (p-1)\frac{(d+1)N_{d+1}(t) - dN_d(t)}{M(t)p} \quad \forall d \geq 2 \tag{16}$$

$$\mathbb{E}(\Delta N_1(t)|(N_0,...)) = -1 + (p-1)\frac{2N_2(t) - N_1(t)}{M(t)p} \tag{17}$$

$$\mathbb{E}(\Delta N_0(t)|(N_0,...)) = 1 + (p-1)\frac{N_1(t)}{M(t)p} \tag{18}$$

Il s'agit donc de montrer que la loi uniforme conditionnellement aux degrés vérifie une certaine "indépendance locale des degrés", c'est-à-dire que dans une hyperarête, si on numérote les sommets de 1 à p, la loi du degré du sommet i (i>1) conditionnellement au degré du sommet 1, est à un terme correctif près la loi du degré d'un sommet dans un hyperarête prise au hasard.
Voici un argument permettant de prouver cette indépendance locale des degrés : l'idée est de partir de la distribution uniforme conditionnellement à la distribution des degrés et à ce que le degré du sommet 1 de l'arête 1 soit d (on suppose les hyperarêtes et les sommets de l'hypergraphe numérotés, par exemple en tirant aléatoirement une numérotation après avoir tiré l'hypergraphe). On applique alors une transformation alétoire à l'hypergraphe pour obtenir un autre hypergraphe, sans changer les degrés des sommets. La transformation est tout simplement de choisir au hasard une hyperarête a ne contenant aucun sommet de l'hyperarête 1 (uniformément parmi celles-ci), puis un sommet x dans cette hyperarête. On intervertit alors leurs places : le sommet x prend la place du sommet 2 dans l'arête 1, et le sommet 2 celle de x dans l'arête a. L'hypergraphe aléatoire G' après cette transformation conserve la même distribution de degrés (et même aucun sommet ne change de degré), toutefois sa loi n'est plus uniforme. Il s'agit de montrer deux choses sur cette transformation : tout d'abord, d'estimer la loi du degré du sommet qui a remplacé 2 dans l'arête 1, puis de montrer que la loi de G' reste proche de la loi uniforme.

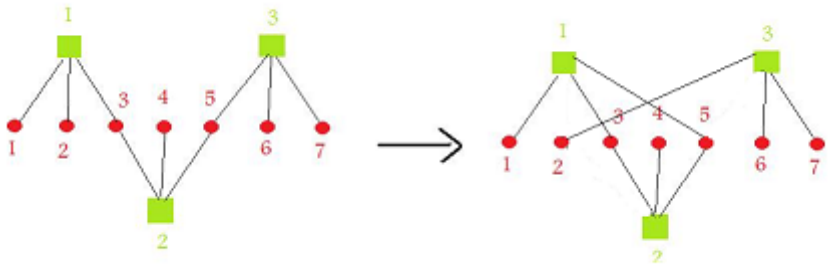

FIGURE 3 – On choisit une arête au hasard ne contenant pas de sommet de l'arête 1 (en l'occurence l'arête 3) et un sommet de cette arête (ici le sommmet 5). le sommet 2 prend la place du sommet 5 dans l'arête 3, le sommet prend la place du sommet 2 dans l'arête 1. Les degrés des sommets sont inchangés

- Etape 1 : Loi du degré après la transformation

L'arête a est choisie parmi les arêtes ne contenant pas un sommet de l'arête 1. Si les degrés ne sont pas trop grands, la proportion de sommets de degré d dans ces arêtes est à peu près la même que dans une arête prise au hasard :

Supposons que le maximum des degrés dans l'hypergraphe est <log(N). Cette hypothèse ne sera pas restrictive pour appliquer le théorème, puisque d'après le lemme 2.4, a.p.s. dans l'état initial tous les sommets sont de degrés $< \log(N)$. Alors il y a moins de $p\log(N)$ sommets appartenant à au moins une hyperarête contenant un sommet de l'hyperarête 1, tous de degré $< \log(N)$. La probabilité pour qu'un d'entre eux soit choisi pour remplacer 2 est donc un $O(\frac{\log(N)^2}{M})$. Un des autres sommets (notons le y) est choisi pour remplacer 2 avec probabilité $\frac{deg(y)}{(M-m)p}$ où m est le nombre d'arêtes contenant un sommet de l'arête 1 ($m \leq p\ \log(N)$). En conclusion, la probabilité pourque le sommet remplaçant 2 soit de degré d est $\frac{d(N_d - O(\log(N))}{(M-O(\log(N))p} + O(\frac{\log(N)^2}{M}) = \frac{dN_d}{Mp} + O(\frac{\log(N)^2}{M})$. Le $O(\frac{\log(N)^2}{M})$ est de plus uniforme pour d entre 0 et log(N) (i.e. l'erreur est majorée par une même fonction $\lambda = O(\frac{\log(N)^2}{M})$ pour tout d).

- Etape 2 : Montrons que la loi de l'hypergraphe G après la transformation est proche de la loi uniforme sur les hypergraphes de distribution de degré $\mathbf{d} = (N_0, N_1...)$ fixée. Plus précisément, notons $\mathbb{P}$ la loi de G après la transformation, et $\mathbb{U}_\mathbf{d}$ la loi uniforme. On va majorer la mesure totale de $\mathbb{U}_\mathbf{d} - \mathbb{P}$. Considérons un hypergraphe G. De combien de G' peut-il provenir en appliquant la transformation ? Autant que de graphes auquel on peut aboutir depuis G en appliquant la transformation : celle-ci est une simple permutation de deux sommets dans deux hyperarêtes, on peut donc obtenir G à partir de G' si et seulement si on peut obtenir G' à partir de G.

On peut donc obtenir G à partir de (M-m)p hypergraphes différents. Soient G' un de ces hypergraphes et m' le nombre d'arêtes contenant un sommet de l'arête 1 dans G'. On peut

obtenir $(M-m^{'})p = Mp + O(\log(N))$ hypergraphes à partir de G' (le O est de plus uniforme en G'), tous avec la même probabilité. Soit K de sorte que pour tout hypergraphe G de distribution de degré , $\mathbb{U}_{\mathbf{d}}(G) = \frac{1}{K}$. Alors on a $\mathbb{P}(G) = \frac{1}{K}\frac{(M+O(\log(N))p}{Mp+O(\log(N))} = \frac{1}{K}(1+O(\frac{\log(N)}{M})$ (avec le O uniforme en G). En sommant sur tous les hypergraphes de distribution de degré $\mathbf{d}$, on obtient $|\mathbb{U}_{\mathbf{d}}-\mathbb{P}| = O(\frac{\log(N)}{M})$ (où on note $|\mu|$ la mesure totale d'une mesure signée μ). Ceci signifie que pour tout évènement A, $|\mathbb{U}_{\mathbf{d}}(A) - \mathbb{P}(A)| = O(\frac{\log(N)}{M})$. On a donc $\mathbb{U}_{\mathbf{d}}(deg(2) = d) = \frac{dN_d}{Mp} + O(\frac{\log(N)^2}{M})$, le O étant uniforme en d.

Conclusion : on se place sur l'évènement degré(x) $<$ log(N) pour tout x dans G(0) (qui a lieu a.p.s. d'après le lemme 2.4). Alors au cours de l'algorithme d'effeuillage, on a, pour une fonction $\lambda = O(\frac{\log(N)^2}{M})$

$$\mathbb{E}(\Delta N_d(t)|(N_0,...)) = (p-1)\frac{(d+1)N_{d+1}(t) - dN_d(t)}{M(t)p} + O(\lambda) \quad \forall d \geq 2 \tag{19}$$

$$\mathbb{E}(\Delta N_1(t)|(N_0,...)) = -1 + (p-1)\frac{2N_2(t) - N_1(t)}{M(t)p} + O(\lambda) \tag{20}$$

$$\mathbb{E}(\Delta N_0(t)|(N_0,...)) = 1 + (p-1)\frac{N_1(t)}{M(t)p} + O(\lambda) \tag{21}$$

Nous pouvons à présent appliquer le théorème 2.1. Il y a une dernière difficulté : on a ici un système infini d'équations différentielles. Cependant, si on tronque à la k-ème équation, en notant $b_k(t) := (p-1)\frac{(k+1)N_{k+1}(t)}{M(t)p}$, fonction bornée par p-1, et T_ϵ=inf {t tq $M < \epsilon N$ ou $N_1(t) = 0$ } alors :

Pour tout $t < T_\epsilon$, on a, avec probabilité $1 - O(\log(N)\lambda\exp(-\frac{N}{\lambda^3})) = 1 - o(1)$:

$$N_d(t) = NY_d(\frac{t}{N}) + O(log(N)^2)$$

, où $Y_d(t)$ est solution du système d'équations différentielles tronqué aux k premières (on note $m(t) = \frac{M(Nt)}{N} = \frac{M(0)}{N} - t$ (à chaque étape de l'algorithme d'effeuillage on supprime une hyperarête)) :

$$Y_d^{'}(t) = (p-1)\frac{(d+1)Y_{d+1}(t) - dY_d(t)}{m(t)p} \; et \; Y_d(0) = \frac{N_d(0)}{N} \quad \forall d \leq k-1$$

$$Y_1^{'}(t) = -1 + (p-1)\frac{2Y_2(t) - Y_1(t)}{m(t)p} \; et \; Y_1(0) = \frac{N_1(0)}{N}$$

$$Y_0^{'}(t) = 1 + (p-1)\frac{Y_1(t)}{m(t)p} \; et \; Y_0(0) = \frac{N_0(0)}{N}$$

$$Y_k^{'}(t) = -(p-1)\frac{kY_k(t)}{m(t)p} + b_k(t) \; et \; Y_k(0) = \frac{N_k(0)}{N}$$

Le fait que la fonction b_k soit borné par p-1 implique que pour k tendant vers l'infini, les solutions Y_i seront proches des solutions du système infini. De plus, puisque les conditions initiales se concentrent autour de $\frac{N_d}{N} = e^{-p\alpha}\frac{(p\alpha)^d}{d!}$, par la continuité des solutions par rapport aux conditions initiales, les $Y_d(t)$ se concentrent autour des solutions du système infini avec conditions initiales $Y_d(0) = e^{-p\alpha}\frac{(p\alpha)^d}{d!}$.

Il nous reste à trouver la solution du système infini. On s'aperçoit en fait qu'on peut chercher une solution qui reste poissonienne pour $d \geq 2$, de paramètre $\gamma(t)$, i.e. $Y_d(t) = e^{-\gamma(t)}\frac{\gamma(t)^d}{d!}$ pour $d \geq 2$. En reportant dans une des équations $d \geq 2$, on doit alors avoir :

$$\gamma^{'}(t) = -\frac{p-1}{\alpha - t}\gamma(t) \; et \; \gamma(0) = p\alpha$$

On en déduit $\gamma(t) = p\alpha(\frac{\alpha - t}{\alpha})^{\frac{p-1}{p}}$. En reportant dans l'équation sur Y_1 puis sur Y_0 on obtient deux équations d'ordre 1 non homogènes, dont les solutions sont :

$$Y_1(t) = \gamma(t)[e^{-\gamma(t)} - 1 + \frac{\gamma(t)}{p\alpha}^{\frac{1}{p-1}}]$$

$$Y_0(t) = 1 - \sum_{k=1}^{\infty} Y_k(t)$$

L'algorithme s'arrête, on le rappelle, quand $N_1 = 0$. Il y a donc deux situations. Pour α petit, la seule solution de $\gamma(t)[e^{-\gamma(t)} - 1 + (\frac{\gamma(t)}{p\alpha})^{\frac{1}{p-1}}] = 0$ est 0, et alors pour N assez grand, l'algorithme d'effeuillage continue au moins jusqu'à ce que le nombre d'arêtes devienne $\leq \epsilon N$, et ce pour tout ϵ. La taille du core est donc alors un o(N). A partir d'une valeur critique $\alpha = \alpha_d$, il existe une solution γ^* de l'équation précédente strictement positive, atteinte en $t = t^*$, alors l'algorithme d'effeuillage s'arrête en laissant un nombre extensif (i.e. proportionnel à N, à un terme correctif près) d'arêtes et de sommets. Pour p=3 par exemple on a $\alpha_d = \approx 0.818469$.

On conclut que les nombres de sommets et d'arêtes du core N_c et M_c sont donnés (à un terme d'erreur o(N) près avec probabilité 1-o(1) d'après l'étude précédente) par :

$$N_c = N\sum_{k=2}^{\infty} Y_k(t^*) = N[1 - (1+\gamma^*)e^{-\gamma^*}] \tag{22}$$

$$M_c = N(\alpha - t^*) = N(\frac{\gamma^*}{p})^{\frac{p}{p-1}}\frac{1}{\alpha^{\frac{1}{p-1}}} = N\alpha(1-e^{-\gamma^*})^p = N\frac{\gamma^*}{p}(1-e^{-\gamma^*}) \tag{23}$$

De plus, la distribution des degrés dans le core est une poissonnienne tronqué aux degré

supérieurs à 2, de paramètre γ^*.

On remarque que contrairement à la transition de percolation, la proportion de sommets est discontinue à la transition.
On peut conclure cette étude par une dernière propriété utile pour la suite : le core, lorsqu'il n'est pas de taille extensive, est vide avec haute probabilité ; plus exactement :

Proposition 2.6. Il existe δ tel que a.p.s. le core de G est soit vide, soit a plus de δN sommets.

Preuve. On pourra consulter dans [6] le lemme 5.1. pour une preuve plus détaillée dans un cadre légèrement différent mais similaire. Soit une partie A de l'ensemble des sommets à s éléments. Si A est l'ensemble des sommets du core de G, alors il a plus de $\frac{2s}{p}$ arêtes. La loi du nombre d'arêtes incluses dans A est $Bin(Binsp, \frac{\alpha p!}{N^{p-1}})$. La probabilité p_s qu'il existe une partie à s éléments contenant plus de $\frac{2s}{p}$ arêtes vérifie donc :

$$p_s \leq \mathbb{P}(Binom(\binom{s}{p}, \frac{\alpha p!}{N^{p-1}}) \geq \frac{2s}{p}) \leq \binom{N}{s} \sum_{k=\frac{2s}{p}}^{\binom{s}{p}} \binom{\binom{s}{p}}{k} (\frac{\alpha p!}{N^{p-1}})^k$$

Dans cette somme, le rapport de deux termes $\binom{\binom{s}{p}}{k}(\frac{\alpha p!}{N^{p-1}})^k$ consécutifs est inférieur à $\frac{p\binom{s}{p}}{2s}\frac{\alpha p!}{N^{p-1}} \leq \frac{p\alpha\delta^{p-1}}{2}$ et ce pour tout $s \in [1, \delta N]$. Si δ est tel que $\frac{p\alpha\delta^{p-1}}{2} < 1$ on a, pour tout $s \in [1, \delta N]$, $p_s \leq C\binom{N}{s}\binom{\binom{s}{p}}{\lceil \frac{2s}{p} \rceil}(\frac{\alpha p!}{N^{p-1}})^{\lceil \frac{2s}{p} \rceil}$ où $C = \sum(\frac{p\alpha\delta^{p-1}}{2})^k = \frac{1}{1-\frac{p\alpha\delta^{p-1}}{2}}$.
En utilisant l'inégalité $\binom{N}{k} \leq \frac{N^k}{k!}$ pour les deux coefficients binomiaux et en remarquant que p>2 implique $\frac{2(p-1)}{p} - 1 > 0$, on en déduit après quelques calculs que $\sum_{s=1}^{\delta N} p_s \longrightarrow_{N\to\infty} 0$, ce qui conclut. □

2.3 Transition de satisfiabilité

2.3.1 Principe de la preuve

Dans cette partie, nous allons, à partir de l'étude du core, déterminer $P_{SAT}(\alpha)$. La démarche adoptée est issue de [3], ou [7], partie E. Nous avons déjà vu que le système est satisfiable si et seulement si le système associé au core l'est. Or, d'après les calculs de la partie 2.3, et d'après la proposition 2.6, a.p.s. le core et vide en dessous du seuil de transition du core α_d. On en déduit que le système associé au core est toujours satisfiable, et que $P_{SAT} = 1$ pour $\alpha < \alpha_d$.

Toutefois, le système p-XORSAT subit une transition de satisfiabilité non pas à $\alpha = \alpha_d$ mais pour une valeur $\alpha_c > \alpha_d$. Un core de taille extensif n'est pas forcément incompatible avec la satisfiabilité du système (contrairement au cas p=2, où l'on a vu que l'existence d'une composante connexe géante impliquait un système non satisfiable). Il faut exploiter de façon plus fine les propriétés du core pour étudier la transition de satisfiabilité ; cette étude va se baser sur les inégalités du premier et second moment.

Proposition 2.7. Inégalités du premier et second moment
Soit Z une variable à valeurs dans $\mathbb{N}$. On a alors :

$$\frac{\mathbb{E}(Z)^2}{\mathbb{E}(Z^2)} \leq \mathbb{P}(Z \geq 1) \leq \mathbb{E}(Z) \tag{24}$$

Preuve. L'inégalité de droite provient simplement de l'inégalité de Markov. L'inégalité de gauche résulte de l'inégalité de Cauchy-Schwarz : $\mathbb{E}(Z)^2 = \mathbb{E}(Z\mathbf{1}_{Z\geq 1})^2 \leq \mathbb{E}(Z^2)\mathbb{E}(\mathbf{1}_{Z\geq 1})$. □

L'idée est d'utiliser cette inégalité pour la variable $Z = \mathcal{N}$, nombre de solutions du système, puisque $P_{SAT} = \mathbb{P}(Z \geq 1)$. On pourrait appliquer ces inégalités au système initial avant effeuillage ; le problème est qu'on ne pourrait pas conclure qu'il y a une transition de satisfiabilité ; on pourrait seulement déterminer des zones de satisfiabilité et de non-satisfiabilité, avec un comportement inconnu entre les deux. Une manière plus puissante d'utiliser ces inégalités est de les appliquer au core de l'hypergraphe. Tâchons donc de calculer les premier et second moment du nombre de solutions du core.

2.3.2 Majoration du seuil de satisfiabilité

Proposition 2.8. Soit un système aléatoire à M équations et N inconnues, où seuls les seconds membres sont des variables aléatoires i.i.d. valant 0 ou 1 avec même probabilité. Alors l'espérance du nombre de solutions est 2^{N-M}

Preuve. Le nombre de solutions s'écrit $\mathcal{N} = \sum_{(x\in\{0,1\}^N} \prod \mathbb{1}_{x_{i_1}+\ldots+x_{i_p}=y}$, où le produit se fait sur toutes les équations, ce qui montre le résultat. □

Ce calcul du premier moment permet déjà d'appliquer l'une des deux égalités précédents. Pour illustrer l'idée précédente que ces inégalités sont plus puissantes appliquées au core, regardons aussi ce qu'on obtient en l'appliquant au système initial. On note $\mathcal{N}$ le nombre de solutions du système initial, et $\mathcal{N}_c$ le nombre de solutions dans le core.
D'après les propositions 2.7 et 2.8, on a $P_{SAT} \leq \mathbb{E}(\mathcal{N}) = \mathbb{E}(2^{N-M})$ d'une part, $P_{SAT} \leq \mathbb{E}(\mathcal{N}_c) = \mathbb{E}(2^{N_c-M_c})$ d'autre part. Les variables M, N_c, et M_c sont concentrées respectivement autour de αN, $N[1-(1+\gamma^*)e^{-\gamma^*}]$, et $N\frac{\gamma^*}{p}(1-e^{-\gamma^*})$. $\mathbb{E}(2^{N-M})$ tend vers 0 si $\alpha > 1$ donc $P_{SAT}(\alpha) = 0$ si $\alpha > 1$. L'inégalité pour le core est plus forte : $\mathbb{E}(2^{N_c-M_c})$ tend vers

0 si $[1-(1+\gamma^*)e^{-\gamma^*}] < \frac{\gamma^*}{p}(1-e^{-\gamma^*})$ (avec les notations de la partie 2.3) ; on en déduit une valeur $\alpha_c < 1$ au-dessus de laquelle $P_{SAT}(\alpha) = 0$. Numériquement on a, par exemple, pour $p = 3$, $\alpha_c = 0,917935$.

2.3.3 Minoration du seuil de satisfiabilité

En fait, α_c est exactement le seuil de transition dans p-XORSAT. Pour le voir, il faut minorer $P_{sat}(\alpha)$ à l'aide de l'autre inégalité, qui fait intervenir le second moment du nombre de solutions dans le core. Il s'agit donc d'étudier ce second moment ; on va s'apercevoir que pour $\alpha < \alpha_c$, le nombre de solutions du core se concentre autour de sa valeur moyenne, i.e.

$$\frac{\mathbb{E}(\mathcal{N}_c)^2}{\mathbb{E}(\mathcal{N}_c^2)} \longrightarrow_{N\to\infty} 1$$

Il en découlera alors que $P_{SAT}(\alpha) = 1$ pour $\alpha < \alpha_c$, ce qui prouvera l'existence d'une transition de satisfiabilité.

Montrons d'abord que le calcul du second moment du nombre de solutions du système associé au core se ramène à celui du premier moment du système homogène associé au core, c'est-à-dire le système obtenu à partir du core en remplaçant tous les seconds membres des équations par 0. On a : $\mathbb{E}(\mathcal{N}_c^2) = \mathbb{E}(\sum_{x,x'} \prod \mathbb{1}_{x_{i_1}+\ldots+x_{i_p}=y} \prod \mathbb{1}_{x'_{i_1}+\ldots+x'_{i_p}=y})$, où les produits ont lieu toutes les équations du système, et où x et x' parcourent tous les N-uplets de booléens. Le changement de variable $z = x' - x$ donne :
$\mathbb{E}(\mathcal{N}_c^2) = \mathbb{E}(\sum_x \prod \mathbb{1}_{x_{i_1}+\ldots+x_{i_p}=y} \sum_z \prod \mathbb{1}_{z_{i_1}+\ldots+z_{i_p}=0})$. Finalement : $\mathbb{E}(\mathcal{N}_c^2) = \mathbb{E}(\mathcal{N}_c)\mathbb{E}(\mathcal{N}_{c,h})$, où $\mathcal{N}_{c,h}$ est le nombre de solutions du système homogène associé au core. On est donc ramenés à montrer que $\mathbb{E}(\mathcal{N}_{c,h}) \sim \mathbb{E}(\mathcal{N}_c)$ lorsque $N \to \infty$. On va montrer le résultat suivant.

Théorème 2.9. On a

$$\mathbb{E}(\mathcal{N}_h) \sim \mathbb{E}(\exp(N \sup_{\omega\in[0,v(\alpha)]} \phi(\omega)))$$

avec $v(\alpha) = e^{-\gamma^*}(e^{\gamma^*} - 1 - \gamma^*)$ et

$$\phi(\omega) = -\omega\log(x(\omega)) - \eta(1-e^{-\eta})\log(1+y(\omega)z(\omega)) + \sum_{l\geq 2} e^{-\eta}\frac{\eta^l}{l!}\log(1+x(\omega)y^l(\omega))$$

$$+\frac{\eta}{p}(1-e^{-\eta})\log(\frac{1}{2}((1+z(\omega))^p + (1-z(\omega))^p))$$

où x, y et z sont des fonctions telles que :

$$\omega = \sum_{l\geq 2} e^{-\eta}\frac{\eta^l}{l!}\frac{x(\omega)y^l(\omega)}{1+x(\omega)y^l(\omega)}$$

$$z(\omega) = \frac{\sum_{l\geq 1}(\eta^l/l!)[x(\omega)y^l(\omega)/(1+x(\omega)y^{l+1}(\omega))]}{\sum_{l\geq 1}(\eta^l/l!)[1/(1+x(\omega)y^{l+1}(\omega))]}$$

$$y(\omega) = \frac{(1+z(\omega))^{p-1} - (1-z(\omega))^{p-1}}{(1+z(\omega))^{p-1} + (1-z(\omega))^{p-1}}$$

Plaçons nous dans le système homogène associé au core. On peut le voir comme un graphe ayant deux types de noeuds, ceux qui représentent les équations et ceux qui représentent les variables (il y a une arête entre une variable x_i et une équation E_j si et seulement si x_i apparaît dans l'équation E_j). Un N-uplet de booléens $(x_1, ..., x_N)$ est solution si et seulement si de tout noeud qui représente une équation part un nombre pair d'arêtes vers des booléens égaux à 1. On note de plus $A = Mp$, le nombre total d'arêtes de ce graphe, qui est aussi la somme des degrés de toutes les variables.

Dans un premier temps, on dénombre les solutions dont un nombre fixé w de variables x_i valent 1. En sommant sur toutes les valeurs possibles de w, on s'apercevra que la somme est équivalente à son plus grand terme ce qui donnera l'expression du moment en fonction d'une borne supérieure sur $\omega = \frac{w}{N}$. Le résultat est le suivant :

Lemme 2.10. Le nombre moyen de solutions du système homogène associé au core ayant w variables valant 1 s'écrit :

$$\mathbb{E}(\mathcal{N}_{c,h,w}) = \mathbb{E}(\sum_{E=0}^{A} \frac{E!(A-E)!}{A!} C(\prod_{0\leq i\leq M}(1+XY^i)^{n_i}X^wY^E)C((\sum_{0\leq 2k\leq p}\binom{p}{2k}X^{2k})^M, X^E))$$

où n_i est le nombre de sommets de degré i pour tout i, et où l'expression $C(P(X), X^n)$ désigne le coefficient de degré n du polynôme P (notation analogue pour les polynômes à deux variables).

Preuve. On remarque que la somme des degrés des variables qui valent 1 doit être égale à la somme, sur toutes les équations, des variables valant 1 dans chaque équation. Fixons donc de plus un entier E entre 0 et A, et choisissons les w variables qui valent 1 de sorte que la somme de leurs degrés soit E. Le nombre de choix possibles est le coefficient de X^wY^E dans le polynôme à deux variables $\prod_{0\leq i\leq M}(1+XY^i)^{n_i}$. A présent, pour chaque équation, choisissons le nombre de variables de l'équation qui valent 1. Il faut choisir, M fois de suite, un entier naturel pair et $\leq p$, de sorte à ce que la somme des M entiers choisis soit E. Le nombre de choix possibles est donc, cette fois, le coefficient de X^E dans le polynôme $(\sum_{0\leq 2k\leq p}\binom{p}{2k}X^{2k})^M$. Rassemblons ensuite variables et équations : il y a $A!$ façons de le faire,

dont seulement $E!(A-E)!$ donnent une solution, d'où le résultat. □

Nous cherchons à présent à estimer les deux coefficients de polynômes ci-dessus. L'idée est la suivante : le coefficient $C(P(X), X^n)$ est le résidu en 0 de la fonction $z \mapsto \frac{1}{z^{n+1}}P(z)$; il s'exprime donc comme une intégrale sur un lacet entourant 0. L'intégrale s'écrit ici sous la forme $\int \frac{1}{z} f^N(z)dz$ asymptotiquement presque sûrement, ce qui va permettre d'appliquer la méthode de Laplace pour en déterminer un équivalent lorsque $N \to \infty$. On n'explicite ici ce calcul que pour le coefficient $C((\sum_{0\leq 2k\leq p} \binom{p}{2k} X^{2k})^M, X^E)$, l'autre se traitant de manière analogue. On pose $Q(X) = (\sum_{0\leq 2k\leq p} \binom{p}{2k} X^{2k})^M = [\frac{1}{2}((1+X)^p + (1-X)^p)]^M$. Le coefficient s'écrit :

$$I_N := C(Q(X)^M, X^E) = \frac{1}{2i\pi} \oint \frac{F_N(z)}{z} dz \tag{25}$$

où $F_N : z \mapsto \frac{1}{z^E} Q(z)^{M_c}$.

Expliquons heuristiquement ce qui va se passer. On rappelle que F_N peut se mettre approximativement sous la forme f^N. En conséquence, lorsque $N \to \infty$, la masse de l'intégrale a tendance à être de plus en plus portée par les z voisins des maxima de $|f|$. On peut alors penser que l'intégrale sur le cercle va être équivalente à l'intégrale sur les petits voisinages des maxima. Notons K le maximum de |f|, et z_1, ..., z_m les points (qu'on suppose être en nombre fini) réalisant ce maximum. Les intégrales de f^N sur des voisinages des z_i sont des $O(K^N)$, et pour qu'elles constituent le terme dominant de l'intégrale sur tout le cercle, elles ne doivent pas être trop petites devant K^N, ce qui exige certaines hypothèses sur le comportement de f au voisinage du maximum. Sous de telles hypothèses, on trouve un équivalent de la forme $CN^a K^N$ et on conclut aisément. Le résultat est le suivant :

Lemme 2.11. On a, pour une certaine fonction ϕ :

$$I_N \sim \frac{1}{\pi} \sqrt{\frac{2\pi}{N|\phi''(0)|}} F_N(z_0)$$

Preuve. Recherchons des points critiques de F_N. En calculant F_N', on voit par les valeurs intermédiaires que F_N' est annulée par un point $z_0 \in \mathbb{R}^{+*}$. $-z_0$ est alors aussi un point critique. D'autre part, $F_N(z_0) = F_N(-z_0)$, et enfin, sur le cercle de centre 0 et de rayon z_0, $|F_N|$ n'atteint son maximum qu'en z_0 et $-z_0$. En effet, $F_N(z) = \frac{1}{z^E} Q(z)^{M_c}$; $\frac{1}{z^E}$ est de module constant sur le cercle est Q est à coefficients réels positifs ; il suffit alors de majorer $|Q(z)|$ par inégalité triangulaire pour obtenir $F_N(z_0)$, et d'utiliser le cas d'égalité dans l'inégalité triangulaire pour obtenir $2\arg(z) = 0[\pi]$, soit $z \in \mathbb{R}$. On choisit ce cercle comme contour d'intégration et on le note C_{z_0}.

On va d'abord estimer l'intégrale de $\frac{F_N(z)}{z}$ sur de petits arcs de cercles de C_{z_0} autour de z_0 et de $-z_0$, puis on montrera que les autres contributions à I sont négligeables.
On se donne un réel $\delta \in]0,\pi[$ et on se place sur l'arc de cercle $A_\delta = \{z_0 e^{it}, |t| < \delta\}$. Si δ est assez petit on peut écrire, avec probabilité $1 - o(1)$ quand $N \to \infty$, $F(z) = \exp(Ng(z))$, où g est holomorphe. Soit $\phi : t \mapsto g(z_0 e^{it}) - g(z_0)$. Alors $\phi(0) = \phi'(0) = 0$, et $\phi''(0) \in \mathbb{R}^{-*}$. On pose de plus $I'_N = 2\pi f(z_0)^{-N} \int_{A_\delta} \frac{F_N(z)}{z} dz = \int_{-\delta}^{\delta} \exp(N\phi(\theta)) d\theta$.

On fait le changement de variable $x = \sqrt{N}\theta$, et pour alléger les notations on pose $u = \frac{1}{\sqrt{N}}$. Il vient $I'_N = u \int_{-\infty}^{\infty} \mathbb{1}_{|x| \leq \delta/u} \exp(\phi(ux)/u^2) dx$. On cherche à appliquer le théorème de convergence dominée lorsque u tend vers 0^+. On a que pour tout $x \in \mathbb{R}$, $\mathbb{1}_{|x| \leq \delta/u)} \exp(\phi(ux)/u^2) \to e^{\frac{x^2 \phi''(0)}{2}}$ lorsque $u \to 0^+$. Choisissons δ assez petit pour avoir $k < 0$ tel que $Re(\phi'') \leq k$ sur $[-\delta, \delta]$. La formule de Taylor avec reste intégral donne que $\phi(ux)/u^2 = x^2 \int_0^1 (1 - s)\phi''(uxs) ds$ donc, pour tout x tel que $|x| \leq \frac{\delta}{u}$, on a $Re(\phi(ux)/u^2) \leq x^2 k \int_0^1 (1 - s) ds = k\frac{x2}{2}$. Ainsi on peut dominer $|1_{|x| \leq \delta/u} \exp(\phi(ux)/u^2)|$ par $e^{kx^2/2}$. Il vient donc : $I'_N \sim \sqrt{\frac{2\pi}{N|\phi''(0)|}}$. Donc finalement, pour δ assez petit,

$$\frac{1}{2i\pi} \int_{C_\delta} \frac{F(z)}{z} dz \sim \frac{1}{2i\pi} \sqrt{\frac{2\pi}{N|\phi''(0)|}} f(z_0)^N$$

On obtient bien sûr le même résultat au voisinage de $-z_0$ avec un arc $A'_{\delta'}$, $\delta' > 0$. Il reste à montrer que l'intégrale de $\frac{F_N(z)}{z}$ sur $C_{z_0} - (A_\delta \cup A'_{\delta'})$ est négligeable devant celles sur A_δ et $A'_{\delta'}$. Ceci découle du fait que le maximum de $|F_N|$ sur le cercle tout entier n'est atteint qu'en z_0 et en $-z_0$. Le maximum K de $|F_N|$ sur $C_{z_0} - (A_\delta \cup A'_{\delta'})$ est alors strictement inférieur à $|F_N(z_0|$, et la contribution de $C_{z_0} - (A_\delta \cup A'_{\delta'})$ à l'intégrale est donc $O(K^N)$, donc négligeable devant les contributions de A_δ et $A'_{\delta'}$, ce qui prouve le lemme. □

Preuve du théorème 2.9. Après un calcul analogue pour le coefficient $C(\prod_{0 \leq i \leq M} (1 + XY^i)^{n_i} X^w Y^E)$, on en déduit une expression de chaque terme de la somme définissant $\mathbb{E}(\mathcal{N}_{c,h,w})$. On observe que ces termes croissent tous exponentiellement en N alors qu'ils ne sont qu'au nombre de N. $\mathbb{E}(\mathcal{N}_{c,h,w})$ est alors équivalent au plus grand de ces termes multiplié par le nombre de valeurs de E qui réalisent le maximum. La valeur de E qui réalise le maximum se calcule facilement. On obtient alors une expression de $\mathbb{E}(\mathcal{N}_{c,h,w})$ faisant intervenir z_0 ainsi que des complexes x_0 et y_0 (qui sont en fait des réels positifs). Notons que z_0, x_0 et y_0 sont tous reliés à w par des conditions qui expriment l'annulation de certaines dérivées. La somme sur w est aussi équivalente à son plus grand terme, on la reparamètre par $\omega = \frac{w}{N}$. En utilisant les propriétés du core trouvées en 2.2.3, on en

déduit finalement l'expression du théorème 2.9. □.

Il ne reste alors qu'à calculer le sup de ϕ. Un calcul direct montre que la dérivée de ϕ s'écrit simplement $\phi'(\omega) = -log(x)$. Pour $\omega = \frac{1}{2}e^{-\gamma^*}(e^{\gamma^*} - 1 - \gamma^*)$, on a $x = y = z = 1$, donc en particulier la dérivée s'annule, et de plus $\phi(\omega) = \log(2)(N_c - M_c)$: c'est l'expression du sup voulue pour montrer l'équivalent $\mathbb{E}(\mathcal{N}_{c,h}) \sim \mathbb{E}(\mathcal{N}_c)$. Pour montrer que cette valeur de ω réalise effectivement le maximum de ϕ lorsque $\alpha < \alpha_c$, il faut montrer que le seul autre maximum local de ϕ est 0, atteint en 0. On admet ici ce dernier point. Notons simplement qu'on a $\phi(0) = 0$ donc, comme $\log(2)(N_c - M_c) < 0$ pour $\alpha > \alpha_c$, on ne peut effectivement espérer montrer que $\log(2)(N_c - M_c)$ est le maximum de ϕ que pour des valeurs $\alpha < \alpha_c$. Énonçons enfin le résultat final :

Théorème 2.12. La probabilité de satisfiabilité lorsque $N \to \infty$ est donnée par :

$$P_{SAT}(\alpha) = \begin{cases} 1 & \text{si } \alpha < \alpha_c, \\ 0 & \text{si } \alpha > \alpha_c \end{cases}$$

P_{SAT} présente donc une discontinuité à la transition, et ne prend même que les valeurs 1 et 0, contrairement au cas p = 2 où P_{SAT} était continue. Comme nous l'avons expliqué au début de cette partie, on a de plus que, contrairement au cas p = 2, la transition de satisfiabilité ne correspond pas à l'apparition d'un phénomène de percolation et on observe en fait deux transitions. Le paragraphe suivant conclut l'étude en décrivant l'autre transition.

2.4 Transition concernant la structure de l'ensemble des solutions

On munit $\{0,1\}^N$ de la distance dite de Hamming définie par $d(x,y)$ égal au nombre d'entiers i tels que $x_i \neq y_i$. On va étudier la géométrie de l'ensemble des solutions du système associé au core pour cette distance. On va constater une transition de phase au voisinage du seuil α_d défini dans la partie 2.2.3. Ce seuil correspond, rappelons-le, à une transition dans la durée de l'algorithme d'effeuillage et donc dans la taille du core. Pour $\alpha < \alpha_d$, les solutions forment un unique "amas" (on va préciser cette notion), tandis que pour $\alpha_d < \alpha < \alpha_c$, elles se répartissent un grand nombre de amas éloignés les uns des autres. Notons que cette transition a lieu alors que sur tout l'intervalle $[0, \alpha_c[$, P_{SAT} reste constant et le nombre moyen de solutions reste continu.

On dit que deux solutions x et y sont dans le même amas si l'on peut passer de x à y par petits "sauts" successifs, où l'on change à chaque "saut" un petit nombre de spins tout en restant dans l'ensemble des solutions. Plus précisemment, on change à chaque étape au plus a_N spins, où a_N est une suite vérifiant $a_N = o(N)$. On peut montrer que deux solutions dans le même amas coïncident nécessairement sur le core :

Proposition 2.13 Supposons $\alpha_d < \alpha < \alpha_c$. Alors il existe $\delta > 0$ tel que, pour tout $\epsilon > 0$, pour tout couple (x, y) de solutions du core, on ait a.p.s : soit $d(x, y) \leq \epsilon N$, soit $d(x, y) \geq \delta N$.

Preuve. Soient x et y deux solutions du core. $x - y$ est solution du système homogène associé au core et $d(x, y) = d(x - y, 0)$ est le nombre de composantes de $x - y$ valant 1. D'autre part, pour $\alpha_d < \alpha < \alpha_c$, on a $\phi'(0) < 0$. Comme $\phi(0) = 0$, il existe donc un $\delta > 0$ tel que $\phi(\omega) < 0$ pour tout ω dans $]0, \delta]$. On en déduit que le nombre moyen de solutions du système homogène associé au core ayant un nombre de composantes qui valent 1 compris entre ϵN et δN décroît donc exponentiellement en fonction de N. En effet, l'expression obtenue dans la partie 2.3 pour $\mathbb{E}(\mathcal{N}_{h,c})$ s'adapte si on restreint w à un intervalle : le résultat est similaire mais le sup porte sur un intervalle en ω différent. Ceci implique en particulier que la probabilité qu'une telle solution existe est un o(1) lorsque $N \to \infty$. Ainsi, a.p.s., $d(x - y, 0) \geq N\delta)$ ou $d(x - y, 0) \leq N\epsilon$, ce qui montre le résultat. □

Pour finir cette étude de la répartition des solutions en amas, nous remarquons qu'il est conjecturé qu'à chaque solution sur le core correspond un unique amas (pour peu qu'on fasse un choix judicieux de a_N ; on pense qu'on peut choisir $a_N = \Theta(\log(N))$) ; ainsi si $\alpha < \alpha_d$ l'espace des solutions forme un seul amas, et est bien connecté. Au contraire, pour $\alpha > \alpha_d$, l'espace des solutions se fragmente en un nombre exponentiel d'amas. On peut trouver dans l'appendice de [3] un argument physique en faveur de cette conjecture.

3 Verre de spins à température non nulle et énergie libre

Nous avons pour l'instant étudié la transition de satisfiabilité, ce qui du point de vue du Hamiltonien

$$\mathcal{H} = \sum_{x_{i_1}+\cdots+x_{i_p}=y \ equation} \frac{1}{2}(1-(-1)^y \sigma_{i_1}\cdots\sigma_{i_p})$$

correspond à la recherche d'états d'énergie nulle. La recherche d'états d'énergie nulle permet de décrire la structure du système à température nulle (où la mesure de Gibbs se concentre dans les états d'énergie minimale), mais pas à température positive.

Dans l'étude du système à température positive, on aimerait comprendre ce qu'il advient des deux transitions vues dans l'étude de p-XORSAT : quel sens leur donner ? Ont-elles lieu pour les mêmes valeurs critiques α_c et α_d qu'à température nulle ?
Pour commencer, on peut définir la deuxième transition, celle qui correspond à la transition de satisfiabilité à température nulle à partir de l'énergie libre. En effet, à température nulle et pour $\alpha < \alpha_c$, la fonction de partition canonique $Z = \sum \exp(-\beta E_i)$ est égale au nombre d'états d'énergie nulle (i.e. de solutions au système). Or on a vu que a.p.s. ce nombre est 2^{N-M} pour $\alpha < \alpha_c$ (et 0 sinon). $\frac{\ln(Z)}{N}$ tend donc en probabilité vers $(1-\alpha)ln(2)$ pour $\alpha < \alpha_c$ et l'énergie libre par spin $F = -\frac{kT\ln(Z)}{N}$ tend en probabilité vers 0. Pour $\alpha > \alpha_c$ et $T \to 0$ en revanche, on a $\frac{ln(Z)}{N} \to -\beta E_{Gs}$ où E_{Gs} est l'énergie d'un état fondamental qui est >0. (i.e. le nombre minimum d'équations non satisfaites par un assignement des N variables divisé par le nombre de spins N). F tend alors vers E_{Gs}. On pourra donc définir la deuxième transition par une discontinuité de l'énergie libre par spins.
L'autre transition ("transition dynamique") dans p-XORSAT, $p \geq 3$, correspond, rappelons-le, non seulement à l'apparition d'un coeur de taille extensive, mais aussi à la transition d'une phase où toutes les solutions sont bien interconnectés, à une phase où elle se concentre dans un grand nombre d'"amas". Il n'existe pas de description aussi simple à température finie, mais on peut envisager plusieurs autres définitions :

On peut caractériser cette transition à partir d'une certaine "longueur de corrélation". Soit $\sigma^{(0)}$ une configuration de spins choisie selon la distribution de Boltzmann de température T. Soit $i \in \{1\cdots N\}$ et l un entier, on note $\mathcal{G}_l(i)$ le sous-graphe contenant tous les sommets de distance strictement inférieure à l de i. On introduit une probabilité sur les configurations de spins :

$$P_{\beta,l} = \frac{exp(-\beta\mathcal{H}(\sigma))}{Z_{\beta,l}} \ si \ \forall j \notin \mathcal{G}_l(i) \ \ \sigma_j = \sigma_j^{(0)} \quad = 0 \ \ sinon \tag{26}$$

et on note $\langle \ \ \rangle_l$ la moyenne par rapport à cette probabilité. $\langle\sigma_i\rangle_l$ est alors correlé à σ_i mais on s'attend à ce que la corrélation diminue lorsque l augmente. On définit alors une

longueur caractéristique :

$$l_i(\epsilon) = \min\{l / \mathbb{E}_{\sigma^{(0)}}(\sigma_i \langle \sigma_i \rangle_l) \leq \epsilon\} \tag{27}$$

Ces longueurs dépendent d'un paramètre $\epsilon \in [0;1]$ qui peut sembler arbitraire, mais ce choix n'a pas d'importance pour définir la transition : celle-ci se caractérise par le fait que tous les $l_i(\epsilon)$ divergent pour T proche de $T_d(\alpha)$ (ou pour α proche de $\alpha_d(T)$ suivant le paramètre qu'on veut faire varier) ; plus exactement : avec probabilité qui ne tend pas vers 0 : $l_i(\epsilon) \sim cste * (T - T_d(\alpha)^{-\frac{1}{2}}$ pour $T > T_d(\alpha)$ ou $l_i(\epsilon) \sim cste * (\alpha_d(T - \alpha)^{-\frac{1}{2})}$ pour α inférieur à $\alpha_d(T)$. On peut vérifier que cette définition de $\alpha_d(T)$ donne la même valeur critique à température nulle que la définition en terme de core du système et de nombre d'amas dans l'ensemble des solutions.
Remarquons tout d'abord que $\langle \sigma_i \rangle_l$ est alors une moyenne uniforme de x_i sur l'ensemble $Sol_{i,l}$ de toutes les solutions σ du système coïncidant avec la solution $\sigma^{(0)}$ sur le complémentaire de $\mathcal{G}_l(i)$. $\langle \sigma_i \rangle_l$ vaut alors soit $\sigma_i^{(0)}$, soit 0 : en effet, soit toutes les éléments de $Sol_{i,l}$ vérifient $\sigma_i = \sigma_i^{(0)}$ et alors $\langle \sigma_i \rangle_l = \sigma_i$, soit il existe une telle solution y avec $y_i \neq \sigma_i$ et exactement la moitié des éléments de $Sol_{i,l}$ vérifie $\sigma_i = \sigma_i^{(0)}$ ($y - \sigma^{(0)}$ est une solution du système homogène valant 0 identiquement sur le complémentaire de $\mathcal{G}_l(i)$ et vérifiant $y_i - \sigma_i^{(0)} = 1$; ajouter $y - \sigma^{(0)}$ constitue une bijection des éléments de $Sol_{i,l}$ vérifiant $\sigma_i = \sigma_i^{(0)}$ vers ceux vérifiant $\sigma_i \neq \sigma_i^{(0)}$. En particulier, $l_i(\epsilon)$ ne dépend ici pas de $\epsilon \in]0;1[$
Si on introduit la probabilité ϕ_l pour que $\langle \sigma_i \rangle_l = \sigma_i^{(0)}$ alors $\mathbb{E}_{\sigma^{(0)}}(\sigma_i \langle \sigma_i \rangle_l) = \phi_l$. Il est possible de calculer récursivement ϕ_l : d'abord il est clair que $\phi_0 = 1$; ensuite on peut établir une formule de récurrence :

$$\phi_{l+1} = \sum_{n=0}^{\infty} e^{-p\alpha} \frac{(p\alpha)^n}{n!} (1 - (1 - \phi_l)^{p-1})^n = 1 - \exp(-p\alpha\phi_l^{p-1}) \tag{28}$$

Cette équation peut se justifier intuitivement comme suit : $e^{-p\alpha}\frac{(p\alpha)^n}{n!}$ est la probabilité que le degré de i soit n ; si le degré de i est n, la valeur de σ_i est déterminée par la valeur des sommets de distance supérieure à l+1 si pour au moins une des arêtes qui le contient les p-1 autres sommets sont déterminés par la valeur des sommets à distance supérieure à l d'eux. On suppose pour écrire cette équation certains évènements indépendants (par exemple le fait que les voisins soient déterminés par leur voisin à distance l) mais il serait possible de le démontrer rigoureusement en utilisant le fait que l'hypergraphe a, à toute distance finie, asymptotiquement presque sûrement, une forme d'arbre.
On voit alors que ϕ_l tend pour l grand vers le plus grand point fixe de $\phi \in [0;1] \mapsto 1 - \exp(-p\alpha\phi^{p-1})$ ce qui est exactement la proportion de sommets dans le core. En dessous de α_d la longueur de corrélation est finie presque sûrement, au dessus elle est infinie avec une probabilité positive. Cette définition permet donc bien d'étendre la définition de la transition dynamique à température nulle.

La transition dynamique à température positive peut aussi se traduire par un comportement différent de chaînes de Markov associés à la mesure de Gibbs-Boltzmann associée au système, avec une divergence cette fois-ci d'un temps de relaxation au niveau de la transition dynamique. Pour une étude détaillée de ce point, on pourra se référer à [8].

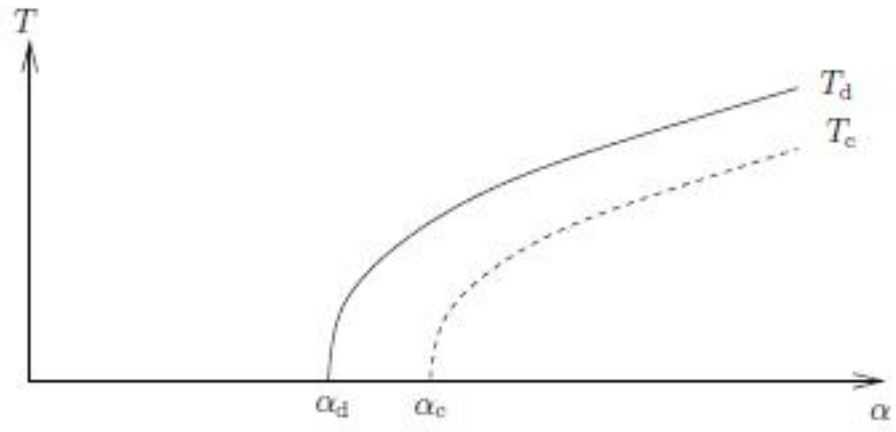

FIGURE 4 – Allure des courbes de transition $\alpha_d(T)$ et $\alpha_d(T)$

Contrairement à l'étude à température nulle, on ne connait pas d'expression des courbes de transition, notamment parce que l'énergie libre est difficile à calculer à toute température et toute connectivité. Il est toutefois assez aisé de calculer l'énergie libre par spins pour des faibles valeurs de la connectivité.

Proposition 3.1 Pour $\alpha < \alpha_c$ et $p \geq 3$, l'énergie libre par spin tend en probabilité vers $F = (1-\alpha)\ln(2) + \alpha\ln(1+e^{-\beta})$

Preuve : Pour $\alpha < \alpha_c$ a.p.s. le système aléatoire est satisfiable. On rappelle que pour $p \geq 3$ notre modèle est de tirer les membres de gauche du système comme le système correspondant à un hypergraphe aléatoire à N sommets où toutes les arêtes sont présentes indépendamment avec probabilité $\frac{\alpha p!}{N^{p-1}}$, et les second membres sont des variables i.i.d. indépendantes des membres de gauche valant 0 ou 1 avec même probabilité.
On peut renoter le système $Ax = b$ (avec $A \in \mathcal{M}_{N,M}(\mathbb{F}_2)$ $x \in \mathbb{F}_2^N$ $b \in \mathbb{F}_2^M$, A,b et également M étant des variables aléatoires). On a alors a.p.s. A est une matrice surjective. En effet, si A n'est pas surjective, $Im(A)$ est un $\mathbb{F}_2$-sev de $\mathbb{F}_2^M$ de codimension au moins 1, d'où comme b conditionnellement à A est un vecteur aléatoire uniforme dans $\mathbb{F}_2^M$, avec probabilité (conditionnellement à A) supérieure à $\frac{1}{2}$, le système n'est pas satisfiable.
Si A est surjective, son noyau est de dimension N-M et contient 2^{N-M} éléments. Il y a alors

2^{N-M} états d'énergie nulle, et pour tout k, $2^{N-M}\binom{N}{k}$ états d'énergie k (il y a exactement $\binom{M}{k}$ vecteurs b' à distance de Hamming k de b, et à chacun de ces vecteurs correspond 2^{N-M} états d'énergie k : les solutions de $Ax = b'$.
La fonction de partition canonique vaut alors

$$Z = \sum_{k=0}^{M} 2^{N-M}\binom{M}{k}e^{-\beta k} = 2^{N-M}(1+e^{-\beta})^M \tag{29}$$

Comme $\frac{M}{N}$ tend en probabilité vers α on en déduit que l'énergie libre par spin tend en probabilité vers $F = (1-\alpha)\ln(2) + \alpha\ln(1+e^{-\beta})$. $\square$.

Remerciements

Nous tenons à remercier chaleureusement Marc Lelarge et Guilhem Semerjian, dont les conseils éclairés et les points de vue complémentaires nous ont été d'une aide précieuse.

Références

[1] Richard Durett, *Random graphs dynamics*, Cambridge University Press, 2006

[2] Rémi Monasson, *Introduction to Phase Transitions in Random Optimization Problems*, Complex Systems, Les Houches Summer School 2006

[3] M. Mézard, R. Zecchina, F. Ricci-Tersenghi, *Alternative solutions to diluted p-spin models and XORSAT problems*,J. Stat. Phys. 111, 505 (2003)

[4] Nick Wormald, *The differential equation method for random graph processes and greedy algorithms*, Lectures on Approximation and Randomized Algorithms, pages 73 155. PWN, Warsaw, 1999.

[5] R.W.R. Darling, J.R. Norris, *Differential equation approximations for Markov chains*, Probability Surveys 2008

[6] Svante Janson, Malwina Luczak, *A simple solution to the k-core problem*, arXiv :math/0508453v1

[7] Marc Mézard, Andrea Montanari, *Information, Physics and Computation*, Oxford University Press, 2009

[8] Guilhem Semerjian, Andrea Montanari, *On the dynamics of the glass transition on Bethe lattices*, J. Stat. Phys. 124, 103 (2006)