THÈSE

présentée
pour obtenir

Le grade de DOCTEUR en SCIENCES
de l'UNIVERSITÉ PARIS-SUD

Spécialité : INFORMATIQUE

par

Yann C. VERHOEVEN

Sujet :

QUELQUES UTILISATIONS DES ARBRES EN COMBINATOIRE

Soutenue devant le jury composé de :

M. Stéphane Boucheron	
M. Olivier Dubois	*Rapporteur*
M. Wenceslas Fernandez de la Vega	*Directeur*
M. Andreas Goerdt	*Rapporteur*
Mme Laurence Puel	*Présidente*
M. Bruce Reed	

Remerciements

Je tiens tout d'abord à remercier Olivier Dubois et Andreas Goerdt d'avoir accepté d'être les rapporteurs de cette thèse. En particulier je dois beaucoup à Olivier pour son soutien infaillible et l'énorme sympathie dont il a fait preuve à mon égard. Je remercie Andreas pour l'inspiration qu'il m'a fournit concernant le problème 2-SAT.

Je remercie ensuite Laurence Puel pour la présidence du jury qui n'a pas du être de tout repos. Merci à Bruce Reed pour ses remarques concernant mon travail.

Je remercie tout particulièrement Stéphane Boucheron pour avoir accepté de faire partie du jury, pour l'aide qu'il m'a fournit pour améliorer la rédaction du manuscrit, mais avant tout je lui suis très reconnaissant pour tous les conseils qu'il m'a donné depuis le DEA, tout au long de ma thèse. Il a été pour moi une très grande source d'inspiration et j'ai découvert énormément de nouvelles choses grâce à lui.

Merci beaucoup à Wenceslas Fernandez de la Vega pour avoir accepté d'encadrer mon stage de DEA puis mon travail de thèse et pour les nombreux conseils prodigués durant ces années.

Je remercie Brigitte Chauvin pour son cours sur les arbres de Galton-Watson qui m'a permis de découvrir ces objets particuliers et de les utiliser dans le cadre de 2-SAT.

Je dois beaucoup à Jean-Michel Drappier grâce à qui lors d'un stage j'ai pu découvrir le côté théorique de l'informatique et réaliser l'existence des algorithmes et de la complexité qui leur est associée.

Merci ensuite au L.R.I. et en particuliers à l'équipe Algorithmique et Complexité pour les différents moyens mis à ma disposition et pour l'accueil qui m'a été fait dès le DEA. En particulier Miklös Santha et Michel de Rougemont qui m'ont permis d'apprécier l'analyse d'algorithmes. Et puis aussi à Yannis Mannoussakis pour m'avoir présenté à lalo et pour le travail qu'il a effectué au moment de ma soutenance.

Ensuite je remercie Christoph Dürr pour l'atmosphère de travail et de calme qu'il m'a aidé à établir dans le bureau durant tout le temps qu'il m'a fallu pour rédiger ma thèse.

Je tiens tout particulièrement à remercier tous les thésards de l'équipe Algorithmes et Complexité avec qui j'ai pu avoir de nombreuses et intéressantes discussions sur des sujets aussi divers que variés. Donc merci à David Gross, Frédéric Magniez, Fabrice Noilhan, Grégory Olocco, Sylvain Peyronnet, Kavé Salamatian et Julien Stern. Ils sont tous devenus beaucoup plus que des collègues.

De même je remercie tout particulièrement Jean-Pierre Tillich pour son soutien exceptionnel tout au long de ma thèse, que ce soit pour la recherche ou pour l'enseignement. Mais surtout merci pour cette amitié.

Je remercie mes parents pour leur aide et leur soutien durant toutes mes études et pour tout ce qu'ils ont su m'apporter et que je ne suis pas capable d'exprimer.

Merci à mon frère Yves avec qui j'ai partagé, je partage et je partagerai des moments exceptionnels, des discussions enrichissantes et j'espère beaucoup d'autres choses.

Enfin je remercie Valéry pour son soutien tout au long de cette thèse et pour la vie heureuse que je partage avec elle.

Et puis je dédie cette thèse à Laury.

Table des matières

Introduction

Aux alentours de 1960, Erdös et Rényi[**15, 16, 17**] ont mis en évidence un comportement particuliers des graphes aléatoires non orientés. En effet, ils ont montré que que certaines propriétés monotones de graphes apparaissent soudainement dans ces graphes aléatoires. Par exemple, on considère l'étude des composantes connexes dans un graphe à n sommets ne contenant aucune arête au début et auquel on ajoute les arêtes une à une au fur et à mesure de l'expérience. Chaque arête est ajoutée à un emplacement tiré aléatoirement parmi les emplacements libres du graphe. On commence avec un graphe vide ne contenant aucune arête et donc aucune composante connexe et on finit avec le graphe complet ayant $n(n-1)/2$ arêtes et qui contient une unique composante connexe. Erdös et Rényi [**16, 17**] ont montré que tant que le nombre d'arêtes ajoutées est inférieur à $n/2$, le graphe contient plusieurs composantes connexes, chacune contenant environ $O(\log n)$ sommets. Par contre, dès que plus de $n/2$ arêtes ont été ajoutées, on constate l'apparition d'une composante de taille εn, où ε dépend du nombre d'arêtes. Cette composante est la seule à être de cette taille, elle est souvent appelée *composante géante du graphe*. On dit alors que l'on a affaire à un *phénomène de seuil*.

Ce comportement n'est pas particulier aux graphes et concerne d'autres structures aléatoires comme nous le verrons dans la suite. De tels comportements d'apparition soudaine d'une propriété ont été aussi observés dans des modèles de mécaniques statistiques et en particulier dans les systèmes désordonnés frustrés lors du passage d'une phase à une autre d'un tel système. C'est pour cela que le terme de *transition de phase* est aussi utilisé pour désigner un tel phénomène.

Lorsque l'on observe un tel phénomène, on se pose alors naturellement la question de savoir à quel moment celui-ci se produit et quelle est la vitesse de ce phénomène. À ce propos, Friedgut [**18**] a montré que pour les propriétés monotones l'apparition de la propriété est très soudain, on parle alors de *seuil raide* (*sharp threshold*).

Ces problèmes, appelés *problèmes de seuil*, ont acquis une certaine importance en informatique en particulier dans l'étude des problèmes **NP**-complets. Nous allons nous intéresser dans ce rapport à deux problèmes présentant des phénomènes de seuil, le problème 2-SAT et la taille de la composante géante fortement connexe dans le graphe aléatoire orienté.

1. Le problème 2-SAT

Considérons le problème diplomatique suivant, tel qu'il a été proposé par Brian Hayes [**25**] :

> Supposez que vous soyez chef du protocole lors du bal de l'ambassade. Le Prince régnant vous demande d'inviter le Pérou, où d'exclure le Quatar. La Reine vous demande d'inviter soit le Quatar, soit la Roumanie, ou les deux. Le Roi, de façon rancunière, souhaite snober soit la Roumanie ou le Pérou, soit les

deux. Existe-t-il une liste d'invités qui satisfasse les caprices de la famille royale ?

Si l'on souhaite exprimer ce problème sous la forme d'une formule simple, on peut le faire de la manière suivante,

$$(P \text{ OU } \neg Q) \text{ ET } (Q \text{ OU } R) \text{ ET } (\neg R \text{ OU } \neg P),$$

où P, Q et R sont des variables booléennes qui ne peuvent prendre que les valeurs **vrai** ou **faux** et ici seront vraies si et seulement si nous invitons, respectivement, le Pérou, le Quatar et la Roumanie. Le symbole $\neg$ indique la négation de la variable qui le suis, ainsi si P a la valeur vraie, l'expression $\neg P$ sera évaluée à faux. L'opération logique OU est définie de façon à ce que (P OU Q) soit évalué à vrai quand P, ou Q est vrai tandis que (P ET Q) est évalué à vrai seulement si P et Q sont vrai.

Le problème consistant à savoir si une formule telle que celle présentée ci-dessus peut être évaluée à vrai est un problème fondamental en informatique et plus connu sous le nom de **SAT**, pour satisfaisabilité. Sous cette forme c'est un problème d'informatique théorique, mais il apparaît aussi dans la pratique, par exemple, lors de la réalisation de logiciels tels que ceux gérant des emplois du temps. On note que dans le cas général si le problème comporte trop de contraintes on ne sait pas le résoudre dans un temps raisonnable. Il est à remarquer que l'on a "trop de contraintes" lorsque celles-ci dépassent le nombre de 50 environ.

Toutefois, nous pouvons remarquer que dans le cas du problème de l'ambassade, la formule est d'une forme très particulière, chaque partie entre parenthèses (appelé *clause* ne contient que deux variables. Le problème consistant à résoudre une telle formule est nommé **2-SAT** et c'est à celui la que nous allons nous intéresser. Comme nous le verrons dans la suite, il est facile de résoudre ce problème, le temps nécessaire étant directement proportionnel au nombre de littéraux. Ce n'est donc pas la résolution de ce problème qui va nous intéresser mais l'étude de la satisfaisabilité des formules en fonction du rapport entre leur nombre de clauses et le nombre de variables. Il est impossible de générer systématiquement toutes les formules dès que le nombre de variables est grand (il y a alors trop de formules), aussi nous allons fixer un nombre de variables et générer des formules aléatoirement. Pour cela on fabrique toutes les formules possibles avec les variables choisies parmi un nombre fixé et on tire uniformément dans cet ensemble. Si on classe les formules tirées par nombre de littéraux et que pour chaque classe on compte le nombre de formules satisfaisables, on observe alors un phénomène étonnant. En effet, lorsque le nombre de littéraux est inférieur au nombre de variables, presque toutes les formules tirées sont satisfaisables tandis que quand le nombre de littéraux est supérieur au nombre de variables elles sont presque toutes insatisfaisables. Une transition se dessine autour de l'endroit ou le nombre de variables et le nombre de clauses sont égaux. C'est ce phénomène que nous allons étudier.

2. La composante fortement connexe

Nous avons vu ci-dessus que dans un graphe aléatoire non dirigé, lorsque le nombre d'arêtes est suffisamment grand, il apparaît une composante connexe géante unique. Or, un phénomène similaire existe dans le cas d'un graphe orienté. Un graphe est dit *orienté* si les arêtes le constituants sont elles mêmes orientées (on les appelle alors des *arcs*), c'est à dire que les arêtes seront représentées par des couples orientés de sommets et non par des paires. Pour construire un tel graphe aléatoirement, on considère le graphe vide (constitué seulement de sommets, sans aucune arête) et on place dans celui-ci chaque arc possible avec probabilité p. Dans un tel graphe, nous dirons qu'il existe un *chemin orienté* entre deux sommets u et

v s'il existe une suite d'arêtes orientés distinctes permettant d'aller de u à v. De plus u et v sont dit être dans la même *composante fortement connexe* s'il existe un chemin orienté de u vers v et un autre de v vers u.

Karp [**26**] a montré que dans un graphe aléatoire orienté, si la probabilité p est de la forme c/n où c est une constante et n est le nombre d'arêtes on observe l'apparition d'une composante fortement connexe géante. Ainsi, si $c < 1$, il existe plusieurs composantes fortement connexes de petites tailles, tandis que si $c > 1$ une composante fortement connexe nettement plus grande que les autres fait son apparition.

Par ailleurs, Karp a aussi donné une estimation des fluctuations de la taille de la composante fortement connexe dans le cas où $c > 1$. Il est a noter que dans le cas du graphe aléatoire non orienté, il a été prouvé par Barraez, Boucheron et Fernandez de la Vega [**6**] que la taille de la composante connexe géante suit une loi Normale.

3. Contributions de la thèse

Dans cette thèse nous allons, comme mentionné ci-dessus, nous intéresser au problème **2-SAT** puis au problème de la composante géante dans le graphe aléatoire dirigé.

Nous avons tout d'abord tenté une unification du problème **2-SAT** en exposant des preuves de théorèmes concernant les principaux résultats. Ainsi, suite à une introduction dans le chapitre 1 aux différents outils qui seront utilisées dans le cadre des démonstrations, nous présentons sommairement dans le chapitre 2 les différentes approches qui ont été utilisées dans l'étude du problème **2-SAT**.

Dans le chapitre 3 nous verrons que la transition se produit en effet autour du point où le nombre de clauses est égal au nombre de variables fixé (ce point est appelé le *seuil*). Ce résultat est dû à Goerdt [**20**], Chvátal et Reed [**10**] et Fernandez de la Vega [**12**], de façon indépendante. La démonstration présentée ici est celle de Chvátal et Reed.

Ensuite, dans le chapitre 4 nous approximerons le nombre de formules satisfaisables en dessous du seuil c'est à dire quand la formule contient moins de clauses que de variable. Dans ce cas, une formule aléatoire est presque sûrement satisfaisable. Ceci est montré en présentant un résultat de Goerdt [**21**]. Toutefois, la démonstration présentée ici a été légèrement modifiée afin d'obtenir un résultat un peu meilleur. On montrera que l'on peut obtenir un résultat substantiellement meilleur en précisant la preuve de Chvátal et Reed dans le cas sous-critique.

Le chapitre 5 est consacré à l'exposé de notre preuve [**37**] concernant le comportement du seuil dans la partie sur-critique, c'est à dire lorsque la formule contient plus de clause que de variables.

Enfin, nous exposerons brièvement dans l'appendice C la preuve proposée par Bollobás, Borgs, Chayes, Kim et Wilson [**7**] donnant la forme exacte du seuil de **2-SAT**.

Dans une seconde partie nous exposerons des résultats classiques concernant la taille de la composante fortement connexe géante dans un graphe aléatoire dirigé dont la probabilité d'arête est c/n avec $c > 1$ et n le nombre de sommets. Après avoir exposé le calcul classique des fluctuations de cette taille, nous montrerons comment obtenir un résultat sensiblement meilleur grâce à l'aide de certains processus de branchements.

La contribution personnelle de cette thèse consiste d'une part en l'apport d'un résultat original concernant une borne supérieure du seuil de transition de **2-SAT** et d'autre part en l'amélioration de résultats déjà connus.

Certains de ces apports ont été obtenus grâce à l'analyse d'algorithmes à l'aide d'arbres de type Galton-Watson. Ces arbres aléatoires ont une probabilité de reproduction constante. Ainsi, un sommet particulier a des descendants dont le nombre est déterminé par une loi de distribution $\mathcal{L}$. Ensuite, chacun des descendants se reproduit lui aussi et son nombre de descendants est déterminé par la loi de distribution $\mathcal{L}$. À chaque génération, les descendants se reproduisent avec cette même loi de distribution $\mathcal{L}$. Une introduction à ces processus constitue l'appendice A et des études plus complètes concernant ces processus "multiplicatifs" peuvent être trouvés dans les ouvrages de Harris [**24**], Athreya et Ney [**5**], Lyons et Peres [**29**] ou Devroye [**22**] par exemple. Les résultats qui nous serons nécessaires dans la suite sont exposés dans l'appendice A de cette thèse.

Cette structure aléatoire a permis d'analyser des algorithmes visant à déterminer les composantes fortement connexes dans un graphe aléatoire orientés. Ainsi nous avons pu préciser la partie supérieure du seuil de transition des formules **2-SAT** aléatoires en montrant, grâce à des résultats connus concernant les arbres de type Galton-Watson, qu'il existe des composantes fortement connexes de tailles suffisantes dans des graphes liés aux formules d'une façon qui sera précisée dans le chapitre 1. Ce résultat est présenté dans le chapitre 5.

Par ailleurs, dans le chapitre 4 les preuves présentées varient légèrement de celles exposées par Goerdt [**21**] d'une part, et Chvátal et Reed [**10**] d'autre part, et ainsi les résultats obtenus sont légèrement meilleurs.

Enfin, dans le chapitre 6, nous précisons le résultat de Karp [**26**] donnant une estimation des fluctuations de la taille de la composante fortement connexe géante dans un graphe aléatoire orienté. Ce résultat est encore obtenu à l'aide d'arbres aléatoires de type Galton-Watson mais cette fois en décomposant ces arbres en sous-arbres ayant certaines propriétés.

CHAPITRE 1

Notations et problèmes

1. Les formules booléennes

Soit un ensemble fini de variables $X = \{x_1, \ldots, x_n\}$. Ces variables sont dites *booléennes* si leurs seules valeurs possibles sont les valeurs de vérité **vrai** et **faux**. Ces variables peuvent être combinées en utilisant des *connecteurs booléens* tels que $\vee$ (*ou* logique), $\wedge$ (*et* logique) et $\neg$ (*négation* logique), de la même façon qu'en algèbre on utilise $+$, $\times$ et $-$ pour former des expressions arithmétiques avec des variables réelles. On peut alors définir une *formule* (ou *expression booléenne*) de la façon suivante :

Définition 1.1. *Une expression constituée de variables booléennes et de connecteurs booléens est appelée une* formule booléenne *si elle est,*

i. une variable booléenne telle que $x_i, i \in \{1, \ldots, n\}$;
ii. une formule de la forme $\neg\phi_1$, où ϕ_1 est une formule booléenne ;
iii. une formule de la forme $(\phi_1 \vee \phi_2)$, où ϕ_1 et ϕ_2 sont des formules booléennes ;
iv. une formule de la forme $(\phi_1 \wedge \phi_2)$, où ϕ_1 et ϕ_2 sont des formules booléennes.

Dans le cas ii la formule $\neg\phi_1$ est appelée la *négation* de ϕ_1. Dans les cas iii et iv, les formules $(\phi_1 \vee \phi_2)$ et $(\phi_1 \wedge \phi_2)$ sont appelées respectivement la *disjonction* et la *conjonction* de ϕ_1 et ϕ_2. Enfin, pour $i \in \{1, \ldots, n\}$, x_i et $\neg x_i$ sont appelés les *littéraux* associés à la variable x_i.

Suivant les variables contenues dans une formule et la façon dont celles-ci sont combinées, cette formule peut-être vraie ou fausse. Les valeurs d'une formule peuvent alors être définies de la façon suivante :

Définition 1.2. *Un* assignement T *est une application d'un ensemble $X' \subset X$ de variables booléennes vers l'ensemble de valeurs de vérité* $\{$**vrai**, **faux**$\}$. *Soit ϕ une formule booléenne. On définit $X(\phi)$, l'ensemble des variables booléennes apparaissant dans ϕ comme suit,*

i. si ϕ est une variable booléenne x_i, $X(\phi) = \{x_i\}$;
ii. si $\phi = \neg\phi_1$, $X(\phi) = X(\phi_1)$;
iii. si $\phi = (\phi_1 \vee \phi_2)$ ou si $\phi = (\phi_1 \wedge \phi_2)$, $X(\phi) = X(\phi_1) \cup X(\phi_2)$.

Soit T un assignement défini sur un ensemble X' de variables booléennes tel que $X(\phi) \subset X'$. T satisfait ϕ *(que l'on note $T \models \phi$) lorsque,*

*i. si ϕ est une variable booléenne x_i, $T \models \phi$ si $T(x_i) = \{$**vrai**$\}$;*
ii. si $\phi = \neg\phi_1$, $T \models \phi$ si $T \not\models \phi_1$ c'est à dire si T ne satisfait pas ϕ_1 ;
iii. si $\phi = (\phi_1 \vee \phi_2)$, $T \models \phi$ si $T \models \phi_1$ ou $T \models \phi_2$;
iv. si $\phi = (\phi_1 \wedge \phi_2)$, $T \models \phi$ si $T \models \phi_1$ et $T \models \phi_2$.

Un assignement satisfaisant une formule ϕ sera appelé un *modèle*.

Deux formules ϕ_1, ϕ_2 sont dites *équivalentes* (noté $\phi_1 \equiv \phi_2$) si pour tout assignement T défini sur un ensemble X' tel que $X(\phi_1) \subset X'$ et $X(\phi_2) \subset X'$, on a $T \models \phi_1$ si et seulement si $T \models \phi_2$ (il est facile de vérifier que l'on a bien défini ici une relation d'équivalence). Si deux formules booléennes sont équivalentes, elles peuvent être considérées comme deux représentations du même objet et peuvent être interchangées. D'autre part on peut aussi remarquer que les connecteurs booléens sont associatifs, commutatifs et distributifs de façon peu différente des opérateurs arithmétiques. Des notations similaires aux notations algébriques $\sum$ et $\prod$ seront utilisées, ainsi on notera $\bigvee_{i=1}^{n} \phi_i$ pour $(\phi_1 \vee \ldots \vee \phi_n)$ et $\bigwedge_{i=1}^{n} \phi_i$ pour $(\phi_1 \wedge \ldots \wedge \phi_n)$.

Une formule booléenne ϕ est dite être en *forme normale conjonctive* (CNF) si $\phi = \bigwedge_{i=1}^{n} C_i$, avec $n \geq 1$, et chaque C_i est une disjonction de littéraux contenant au moins un littéral. Les C_i sont appelées les *clauses* de l'expression. De façon analogue, une formule ϕ est dite être en *forme normale disjonctive* (DNF) si $\phi = \bigvee_{j=1}^{n} D_j$, avec $n \geq 1$, et chaque D_j est une conjonction contenant au moins un littéral. Les D_j sont appelés les *implicants* de la forme normale disjonctive. On peut alors énoncer le théorème suivant,

Théorème 1.3. *Toute formule booléenne est équivalente à une formule en forme normale conjonctive et à une autre en forme normale disjonctive.*

Une démonstration de ce théorème et des précisions sur ce qui a été dit peuvent être trouvées dans [**33**].

Enfin, deux littéraux u et v sont dits *strictement distincts* si on a ni $u = v$, ni $u = \neg v$.

2. La satisfaisabilité d'une formule

Une formule booléenne ϕ est dite *satisfaisable* s'il existe un assignement T défini sur un ensemble X' tel que $X(\phi) \subset X'$ tel que $T \models \phi$. La formule ϕ est dite être *valide* ou une *tautologie* si pour tout T défini sur un ensemble X' tel que $X(\phi) \subset X'$, $T \models \phi$. On remarque qu'une expression valide doit être satisfaisable. Si ϕ est valide on peut écrire $\models \phi$ sans référence à un assignement T puisque tout assignement satisfait cette formule. Par ailleurs, supposons que ϕ soit une expression booléenne insatisfaisable, alors il est clair que pour tout assignement T défini sur un ensemble X' tel que $X(\phi) \subset X'$, $T \not\models \phi$ et donc $T \models \phi$. On en déduit donc :

Proposition 1.4. *Une expression booléenne est* insatisfaisable *si et seulement si sa négation est valide.*

La satisfaisabilité est une propriété très importante des formules booléennes. Dans la suite nous nous intéresserons aux problèmes algorithmiques liés à cette propriété. En théorie de la complexité, le problème SATISFAISABILITÉ (communément appelé SAT) est très étudié. Traditionnellement, les expressions booléennes utilisées dans les algorithmes sont représentées comme un mot d'un alphabet contenant les symboles $x, 0$ et 1 (pour pouvoir écrire les noms de variables et leurs indices en binaires) et les symboles $(,), \vee, \wedge$ et $\neg$ dont on a besoin pour écrire les expressions booléennes. La *longueur* d'une expression booléenne est alors la taille du mot correspondant. Le problème SAT est le suivant,

SAT

entrée : une formule booléenne en forme normale conjonctive,
sortie : **oui** *si cette formule est satisfaisable,* **non** *sinon.*

On peut noter que l'on exige ici que la formule soit en forme normale conjonctive. Il y a deux raisons à cela. Tout d'abord on sait que toute formule peut être mise sous cette forme et ainsi nous disposons d'une structure commune pour les entrées du problème. Ensuite, cette forme particulière semble plus aisée à manier. Par exemple, considérons la formule

$$\phi = ((x_1 \vee x_2 \vee x_3) \wedge (x_1 \vee \neg x_2) \wedge (x_2 \vee \neg x_3) \wedge (x_3 \vee \neg x_1) \wedge (\neg x_1 \vee \neg x_2 \vee \neg x_3))$$

Cette formule est bien en forme normale conjonctive et donc un assignement doit satisfaire toutes les clauses. La première clause impose que $T(x_1)$, $T(x_2)$ ou $T(x_3)$

soit **vrai**. Les trois clauses suivantes imposent que les trois valeurs $T(x_1)$, $T(x_2)$ et $T(x_3)$ soient égales. Et enfin la dernière clause impose qu'au moins l'une des valeurs $T(x_1)$, $T(x_2)$ ou $T(x_3)$ soit **fausse**. Cette formule est donc insatisfaisable.

Ainsi, une formule en forme normale conjonctive est satisfaite par un assignement si celui-ci satisfait toutes ses clauses.

On peut noter qu'un algorithme exhaustif peut résoudre le problème SAT en temps $O(n2^n)$ en essayant toutes les combinaisons possibles de valeurs de vérité pour les variables apparaissant dans l'expression. D'autre part, un algorithme non déterministe qui devine un assignement satisfaisant et vérifie qu'il satisfait bien toutes les clauses, résout facilement ce problème. On en déduit donc que SAT est dans la classe de complexité **NP**. De plus le théorème de Cook nous dit que SAT est un problème **NP**-complet. Pour une démonstration de ce théorème, on pourra se référer à [**33**].

3. La notation de Landau

Soient les fonctions f et g dans $\mathbb{R}$. Dans la suite nous allons utiliser très souvent la notation de Landau $f = O(g)$ et $f = o(g)$. On rappelle que la notation $f = O(g)$ signifie $|f| \leq c|g|$, où c est une constante. Si nous avons une inégalité de la forme $|f| \leq h(g)|g|$, où h est une fonction majorée, pas nécessairement uniformément, pour g fini et qui est uniformément majorée quand g tend vers 0, on notera alors $f = \mathbb{O}_0(g)$. Par exemple, avec cette notation, $e^{x^2} - 1$ est $\mathbb{O}_0(x^2)$, mais n'est pas $O(x^2)$ (on prend $h : x \mapsto e^x$).

La notation $f = o(g)$ quand à elle signifie que f/g tend vers 0 quand les variables tendent vers leurs limites. De plus la notation $f = o_{\lambda,\varepsilon}(g)$ signifie que $f/g \to 0$ quand $\lambda \to \infty$ et $\varepsilon \to 0$.

Enfin, $f = \Theta(g)$ signifie qu'il existe des constantes finies c_1 et c_2 telles que $c_1 \leq f/g \leq c_2$.

Dans la suite, en l'absence d'indication, la quantité tendant vers une limite dans les expressions utilisant la notation de Landau sera le nombre de variables n des formules considérées.

4. Le problème k-SAT

Le théorème 1.3 nous indique que toute formule booléenne peut s'écrire sous forme normale conjonctive. On peut alors classer les formules selon le nombre de littéraux maximum contenu dans chaque clause. Ainsi, pour $k \geq 1$ fixé, si chaque clause d'une formule F, en forme normale conjonctive, contient k littéraux, on dira que cette formule est k-CNF. Il faut noter que les répétitions de littéraux dans une clause sont autorisées. Ceci permet de simplifier certaines preuves. On peut alors définir le problème k-SAT comme suit :

k-SAT

entrée : une formule booléenne k-CNF,
sortie : **oui** *si cette formule est satisfaisable,* **non** *sinon.*

Il a été montré que le problème 2-SAT appartient à la classe de complexité **P** (voir [**4**]).

Par contre, il en va tout autrement dès que $k \geq 3$, le problème k-SAT est alors **NP**-complet. Il est clair que tout problème du type k-SAT appartient à la classe de complexité **NP** (la démonstration est la même que pour SAT). Par ailleurs, pour $k \geq 3$ fixé, on peut transformer toute formule SAT en formule k-CNF, donc on déduit la **NP**-complètude de k-SAT. Pour plus de précisions sur ces résultats, on pourra consulter [**33**].

Toutefois, nous allons ici nous préoccuper d'un aspect différent du problème k-SAT. Tout d'abord, nous considérons un aspect aléatoire du problème. Soit $k > 1$, choisissons aléatoirement une formule F dans l'ensemble de toutes les formules k-CNF sur n variables, muni d'une distribution uniforme. Soit m le nombre de clauses de F. Alors, on conjecture l'existence pour chaque k d'une constante c_k telle que si $m < c_k n(1 - o(1))$ F est presque sûrement satisfaisable tandis que si $m > c_k n(1 - o(1))$, F est presque sûrement insatisfaisable (une démonstration de ce résultat peut être trouvée dans [**10**]). De plus il a été montré ([**10, 20**]) que $c_2 = 1$, mais on ne connaît pas c_k pour $k > 2$. Des simulations semblent indiquer que $c_3 \approx 4,25$ et [**8, 9, 30, 19, 13, 28, 1, 14**] obtiennent des bornes inférieures et supérieures de c_3. La meilleure borne inférieure est 3,145, due à Achlioptas [**1**] et la meilleure borne supérieure est 4,506, due à Dubois, Boufkhad et Mandler [**14**].

Nous sommes donc ici en présence d'un phénomène de seuil. Il a été montré par Friedgut [**18**] que pour $k > 1$ la transition de part et d'autre du seuil est "dure" mais sans en préciser la pente. Dans la suite nous nous intéresserons au cas $k = 2$ car bien que la valeur de c_2 soit parfaitement connue, plusieurs travaux ([**21, 37, 7**]) ont permis de déterminer avec exactitude la forme du seuil.

5. Des formules booléennes aux graphes

Un outil important qui a permis de traiter le problème de la forme du seuil de 2-SAT est la transformation introduite par Aspvall, Plass et Tarjan [**4**] qui permet d'obtenir un graphe dirigé à partir d'une formule 2-CNF. Cette transformation est intéressante car la satisfaisabilité de la formule est équivalente à une propriété de connexité du graphe obtenu. Il faut toutefois noter qu'il existe d'autres transformations d'une formule 2-SAT en graphe comme celle utilisée par Fernandez de la Vega [**12**]. Nous allons présenter la transformation de Aspvall, Plass et Tarjan car c'est celle qui a été la plus utilisée et c'est celle utilisée dans la suite.

Soit F une formule 2-CNF sur n variables et contenant m clauses (on suppose sans perte de généralité que toute les clauses sont de taille 2, en effet, si une clause de la forme u est présente dans la formule, on la transforme en $u \vee u$ qui est équivalente). Soit $G(F)$ le graphe obtenu à partir de la formule F. La construction est la suivante :

i. pour chaque variable u_i de la formule F, on introduit 2 sommets x_i et $\neg x_i$ au graphe $G(F)$;
ii. pour chaque clause $u_i \vee u_j$ de F, on ajoute les arêtes $\neg x_i \to x_j$ et $\neg x_j \to x_i$ au graphe $G(F)$.

Par exemple considérons la formule $\phi_2 = (u_1 \vee u_2) \wedge (u_1 \vee \neg u_2) \wedge (u_2 \vee u_3) \wedge (\neg u_1 \vee \neg u_3)$ et intéressons nous à son graphe associé. La première clause est $(u_1 \vee u_2)$, les variables sont u_1 et u_2, on introduit donc les sommets x_1, x_2, $\neg x_1$ et $\neg x_2$ dans le graphe avec les arêtes orientées $\neg x_2 \to x_1$ et $\neg x_1 \to x_2$. Le seconde clause $u_1 \vee \neg u_2$ concerne encore les variables u_1 et u_2 donc on introduit pas de sommet, mais les arêtes orientées $\neg x_1 \to \neg x_2$ et $x_2 \to x_1$. La troisième clause $u_2 \vee u_3$ contient la variable u_3, donc on introduit les sommets x_3 et $\neg x_3$ et on a les arcs $\neg x_2 \to x_3$ et $\neg x_3 \to x_2$. Et enfin, la dernière clause $\neg u_1 \vee \neg u_3$ permet d'obtenir les arcs $x_1 \to \neg x_3$ et $x_3 \to \neg x_1$. Le graphe obtenu est le graphe présenté dans la figure 1.

Dans la suite, nous dirons qu'un graphe orienté est *fortement connexe* si pour tout couple de sommets (u, v) il existe un chemin orienté de u vers v et un chemin de v vers u. Les sous-graphes maximaux fortement connexes sont appelés les *composantes connexes fortes* du graphe $G(F)$. Il est clair que de telles composantes ont des ensembles de sommets disjoints. De plus, par construction de $G(F)$, pour chaque composante fortement connexe S de $G(F)$ il existe une composante fortement connexe complémentaire $\neg S$ constituée du sous-graphe induit par les compléments des sommets de S (éventuellement ces deux sous-graphes sont confondus).

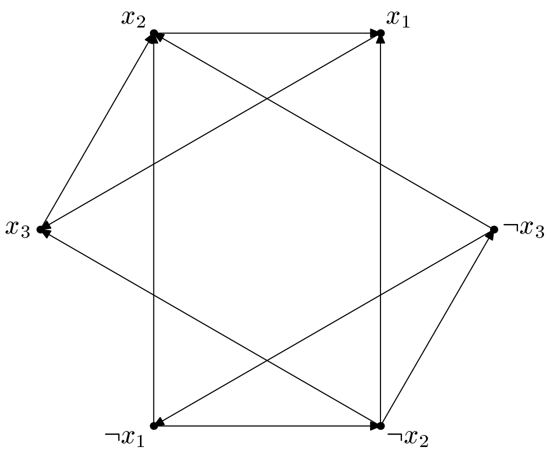

FIGURE 1. graphe associé à la formule $\phi_2 = (u_1 \vee u_2) \wedge (u_1 \vee \neg u_2) \wedge (u_2 \vee u_3) \wedge (\neg u_1 \vee \neg u_3)$

Supposons que l'on assigne des valeurs de vérité aux sommets de $G(F)$. Un tel assignement satisfait F si et seulement si

a. pour tout $i \in \{1, \ldots, n\}$, u_i et $\neg u_i$ reçoivent des valeurs de vérité complémentaires ;
b. il n'existe pas d'arête $u \to v$ telle que u soit **vrai** et v **faux** (c'est à dire qu'aucun chemin ne mène d'un sommet **vrai** à un sommet **faux**).

On peut alors énoncer le théorème suivant,

Théorème 1.5 (Aspvall, Plass et Tarjan, 1979)**.** *Une formule F est satisfaisable si et seulement si dans $G(F)$ aucun sommet u_i n'est dans la même composante fortement connexe que son complémentaire $\neg u_i$ (*i.e.* aucune composante fortement connexe S n'est égale à son complémentaire $\neg S$).*

Preuve du Théorème 1.5 (Aspvall, Plass et Tarjan, 1979). Soit un sommet u_i dans la même composante fortement connexe que $\neg u_i$, alors tout assignement sur les sommets de $G(F)$ contredit a ou b. Inversement, l'algorithme de 2-satisfaisabilité présenté dans la suite détecte la présence d'un littéral et de sa négation dans la même composante fortement connexe ou marque les composantes fortement connexes d'une façon qui implique l'existence d'un assignement satisfaisant F. □

Remarque 5.6. Il est a noter qu'il existe un algorithme permettant de déterminer les composantes fortement connexes d'un graphe dirigé. Cet algorithme, dû à Tarjan [**36**], repose sur le principe de la recherche en profondeur.

Algorithme de 2-satisfaisabilité
Pour chaque composante fortement connexe S de $G(F)$,

1. si S est marquée, ne rien faire ;
2. si $S = \neg S$, stop : F est insatisfaisable ;
3. marquer S **vrai** et $\neg S$ **faux**.

Cet algorithme s'arrête prématurément si un sommet est dans la même composante fortement connexe que son complémentaire indiquant l'insatisfaisabilité. En utilisant la dualité du graphe et une preuve par récurrence sur l'exécution de l'algorithme on montre que chaque composante connexe forte marquée **vrai** n'a que des composantes **vrai** comme successeurs et chaque composante marquée **faux** n'a que des composantes **faux** comme prédécesseurs. Ainsi si l'algorithme ne finit pas prématurément, il marque chaque composante de telle façon que des composantes complémentaires ont des valeurs complémentaires et aucun chemin ne va d'une composante **vrai** à une composante **faux**. Si on assigne à chaque sommet la valeur de la composante fortement connexe qui le contient, on obtient un assignement qui vérifie *a.* et *b.*

Dans la suite, nous appellerons *cycle contradictoire* un cycle contenant un littéral et sa négation dans une composante fortement connexe.

Lorsqu'il réussit, cet algorithme fournit ainsi un assignement satisfaisant la formule F. De plus on peut remarquer que la réponse est donnée en temps $O(n + 2m)$. Cela permet de vérifier que 2-SAT est bien dans **P**.

6. Rappels concernant les graphes aléatoires

Dans la suite, nous allons utiliser la théorie des graphes aléatoires introduite par Erdös et Rényi [**16, 17**].On rappelle que dans ce contexte, une propriété sera dite "presque sûre" si la probabilité que cette propriété se réalise tend vers 1 quand n tend vers l'infini.

Les graphes aléatoires utilisés seront des graphes orientés et non orientés.

6.1. Le modèle non orienté. Soit p tel que $0 \leq p \leq 1$. Considérons un graphe vide (sans arêtes) dont les sommets sont $x_1, \ldots, x_n$. On considère alors toutes les arêtes possibles $x_i x_j$, pour $1 \leq i < j \leq n$, et on ajoute chacune de ces arête dans le graphe avec probabilité p. Ce graphe sera noté $G_{n,p}$. C'est le modèle proposé par Erdös et Rényi.

Dans un graphe, nous dirons qu'il existe un *chemin* entre deux sommets x et y si il existe des sommets $y_0, \ldots, y_k$ tels que $x = y_0$, $y = y_k$ et les arêtes $y_i y_{i+1}$ existent pour $0 \leq i < k$. Nous dirons alors que ce chemin est de longueur $k - 1$.

6.2. Le modèle orienté. Dans la suite, nous appellerons *arcs* les arêtes orientées.

Soit p tel que $0 \leq p \leq 1$. Comme précédemment on considère le graphe vide de sommets $x_1, \ldots, x_n$ et on ajoute les arcs un à un, chacun avec probabilité p. Le graphe obtenu est le graphe aléatoire $D_{n,p}$.

Dans un graphe, nous dirons qu'il existe un *chemin orienté* (ou encore *chemin dirigé* simplement *chemin* lorsqu'il n'y a pas d'ambiguïté) entre deux sommets x et y s'il existe des sommets $y_0, \ldots, y_k$ tels que $x = y_0$, $y = y_k$ et les arcs $y_i y_{i+1}$ existent pour $0 \leq i < k$. Nous dirons alors que ce chemin orienté est de longueur $k - 1$.

De plus, nous dirons qu'un sommet v est *atteignable* (ou *accessible*) à partir du sommet u s'il existe un chemin dirigé allant de u à v.

7. Les modèles de 2-SAT aléatoire

Dans la suite, nous allons nous intéresser au comportement des formules 2-CNF aléatoires. Il nous faut donc définir le modèle probabiliste que nous allons utiliser. Il existe deux modèles qui sont presque équivalents (dans le sens où un résultat dans l'un des deux modèle peut être facilement transposé dans l'autre modèle).

Soit un ensemble de variables booléennes $\{x_1, \ldots, x_n\}$. Dans le premier modèle, on considère l'ensemble constitué de toutes les formules 2-CNF contenant exactement m clauses (la clause $u \vee v$ est la même que la clause $v \vee u$) et on choisit les formules $F_{n,m}$ de manière uniforme dans cet ensemble. C'est souvent en utilisant ce modèle que seront énoncés les théorèmes.

Dans le second modèle, on considère l'ensemble des $4\binom{n}{2}$ clauses différentes possibles avec n variables et chacune de ces clauses est présente avec probabilité p dans la formule $F_{n,p}$.

Pour les démonstrations, c'est le second modèle qui sera utilisé en posant

$$\begin{aligned} p &= \frac{m}{4\binom{n}{2}} \\ &= \frac{m}{2n^2}(1 + O(\frac{1}{n})). \end{aligned}$$

Le lecteur interessé pourra consulter l'appendice A de [**7**] pour une comparaison précise entre les deux modèles et en particulier pourquoi la différence existant est négligeable pour les résultats obtenus ici.

CHAPITRE 2

Survol des techniques de résolution du problème 2-SAT aléatoire

1. Introduction

Nous allons ici évoquer les diverses méthodes qui ont été utilisées pour étudier le problème 2-SAT aléatoire que nous avons présenté dans le chapitre 1. Comme nous allons le voir, la plupart de ces méthodes comportent des points communs.

On va aussi voir que le problème 2-SAT est en fait composé de deux problèmes. Il faut tout d'abord déterminer s'il existe une constante c fixée pour laquelle, pour tout $\varepsilon > 0$, si $p < (c-\varepsilon)/n$ alors la probabilité que $F_{n,p}$ soit satisfaisable tend vers 1 quand n tend vers l'infini, tandis que si $p > (c+\varepsilon)/n$, la probabilité que $F_{n,p}$ soit satisfaisable tend vers 0 quand n tend vers l'infini. Ce problème est discuté dans la section 3 de ce chapitre et dans le chapitre 3 de manière plus précise.

Le second problème est de connaître avec autant de précision que possible la probabilité de satisfaisabilité au seuil et autour du seuil. On cherche à déterminer un intervalle $I(n)$ de longueur maximale tel que, pour tout $\rho = m/n \in I(n)$, la probabilité que $F_{n,m}$ soit satisfaisable et la probabilité que $F_{n,m}$ soit insatisfaisable soient toutes les deux supérieures à une constante $\mu > 0$ quand n tend vers l'infini. Ce problème sera présenté dans les section 4 et 5 de ce chapitre puis traité dans les chapitres 4, 5 et l'appendice C avec précision.

Dans toute l'étude de 2-SAT aléatoire, le graphe proposé par Aspvall, Plass et Tarjan [**4**] (présenté dans la section 5 du chapitre 1) est sans doute la façon de représenter les formules 2-CNF la plus utilisée. En effet, pour montrer qu'une formule n'est pas satisfaisable, il suffit de montrer qu'il existe un littéral et sa négation dans une même composante connexe. Différentes approches visant à montrer ce résultat ont été développées et nous allons les exposer brièvement dans ce chapitre après avoir défini les différentes structures utilisées. Les démonstrations seront pour la plupart faites dans la suite.

Toutefois ce n'est pas la seule façon d'envisager ce problème. En effet, Fernandez de la Vega [**12**] a pu déterminer le seuil en utilisant un autre type de graphe. Nous présenterons cette méthode dans la section 6.

Par ailleurs, une revue des méthodes de résolution du problème de 2-SAT aléatoire a été réalisée par Fernandez de la Vega [**11**].

2. Des structures dans un graphe orienté

2.1. Les bicycles. La notion de *bicycle* a été définie par Chvátal et Reed [**10**] dans le cadre de leur étude de 2-SAT en dessous du seuil. La définition est la suivante,

Définition 2.1. *Nous appelons* bicycle *une formule contenant les variables distinctes* $x_1, x_2, \ldots, x_s$ *(avec* $s \geq 2$*) et les clauses distinctes* $C_0, C_1, \ldots, C_s$ *et telle qu'il existe des littéraux* $l_1, l_2, \ldots, l_s$ *vérifiant le fait que chacun des* l_r *est soit* x_r*, soit* $\neg x_r$ *et chaque clause* C_r*,* $0 < r < s$*, est égale à* $\{\neg l_r, l_{r+1}\}$ *avec* $C_0 = \{u, l_1\}$ *et* $C_s = \{\neg l_s, v\}$*, où* u *et* v *appartiennent à* $\{x_1, \ldots, x_s, \neg x_1, \ldots, \neg x_s\}$.

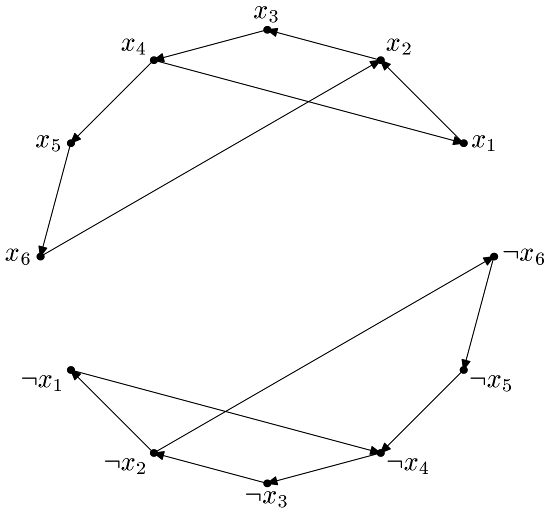

FIGURE 2. La représentation du bicycle Φ_1

D'autre part, nous dirons qu'une formule contient un bicycle de taille r si l'une de ses sous-formules est un bicycle et porte sur r variables.

Par exemple, considérons la formule Φ_1

$$\Phi_1 = (\neg x_4 \vee x_1) \wedge (\neg x_1 \vee x_2) \wedge (\neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4) \wedge (\neg x_4 \vee x_5) \wedge (\neg x_5 \vee x_6) \wedge (\neg x_6 \vee x_2).$$

On peut vérifier facilement que cette formule est un bicycle. Par ailleurs, si on représente cette formule par un graphe comme cela a été évoqué dans la section 5 du chapitre 1, on obtient la figure 2 dans laquelle on peut remarquer deux cycles.

Remarque 2.2. Il faut toutefois noter qu'une formule contenant un cycle ne contient pas nécessairement un bicycle. Ainsi, la formule Φ_2

$$(\neg x_4 \vee x_1) \wedge (\neg x1 \vee x_2) \wedge (\neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4)$$

contient bien un cycle (non contradictoire, la formule est satisfaisable) comme on peut s'en rendre compte sur la figure 3 mais ne contient pas de bicycle. En effet, chaque variable n'apparaît que deux fois dans la formule tandis que dans un bicycle au moins une variable apparaît strictement plus de deux fois.

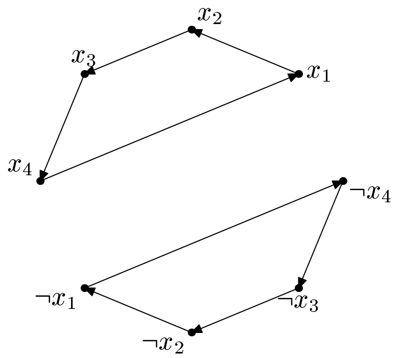

FIGURE 3. La représentation de la formule Φ_2

Dans la suite, nous utiliserons les bicycles et en particulier le lemme suivant,

Lemme 2.3. *Une formule 2-SAT insatisfaisable contient un bicycle.*

Preuve du lemme 2.3. Soit F une formule sous la forme 2-CNF et soit $G(F)$ son graphe associé. Nous avons vu dans la section 5 du chapitre 1 que F est insatisfaisable si et seulement si $G(F)$ contient un circuit contradictoire c'est à dire si $G(F)$ contient un chemin dirigé $w_0, w_1, \ldots, w_i$ tel que $w_0 = w_i$ et, tel que, pour un certain t, $w_t = \neg w_0$. En choisissant un chemin qui minimise t, on trouve que tous les littéraux $w_1, w_2, \ldots, w_t$ constituant ce chemin sont distincts et aucun n'est le complément d'un autre. Si s est le plus grand indice tel que tous les littéraux $w_1, w_2, \ldots, w_s$ soient distincts et qu'aucun d'entre eux ne soit le complément d'un autre, alors les clauses $C_0, C_1, \ldots, C_s$, avec $C_r = \{\neg w_r, w_{r+1}\}$, forment un bicycle. Ainsi nous avons montré qu'une formule insatisfaisable contient un bicycle. □

Remarque 2.4. La réciproque de ce lemme est fausse. Par exemple, la formule Φ_2 est un bicycle, mais est satisfaisable (on note l'absence de cycle contradictoire sur la figure 2.

2.2. Les serpents. Chvátal et Reed ont définis une seconde structure dans un graphe associé à une formule, le *serpent*. Cette structure est définie de la manière suivante,

Définition 2.5. *Un* serpent *est une suite de littéraux distincts* $l_1, l_2, \ldots, l_s$*, avec* $s = 2t - 1$*, telle qu'aucun* l_i *n'est le complément d'un autre.*

En posant $l_0 = l_{s+1} = \neg l_t$, à chaque serpent A, on associe l'ensemble F_A des $s+1$ clauses $\{\neg l_r, l_{r+1}\}$, avec $0 \leq r \leq s$. Si l'on considère F_A comme une formule, on remarque que cette formule est insatisfaisable. Considérons le graphe $G(F_A)$ associé à cette formule. On voit que pour tout $(i,j) \in \{0, s+1\}^2$, $i \leq j$, il existe un chemin dirigé de l_i vers l_j, et ainsi un chemin de $l_0 = \neg l_t$ à l_t et de l_t à $l_{s+1} = \neg l_t$. Donc il existe un cycle contradictoire dans $G(F_A)$.

On peut par exemple considérer la formule Φ_3 définie de la façon suivante,

$$\Phi_3 = (x_3 \vee x_1) \wedge (\neg x_1 \vee x_2) \wedge (\neg x_2 \vee x_3) \wedge (\neg x_3 \vee x_4) \wedge (\neg x_4 \vee x_5) \wedge (\neg x_5 \vee \neg x_3).$$

Alors cette formule est la formule associée au serpent $x_1, \ldots, x_5$. Le graphe associé à ce serpent est celui représenté dans la figure 4. On voit le cycle contradictoire

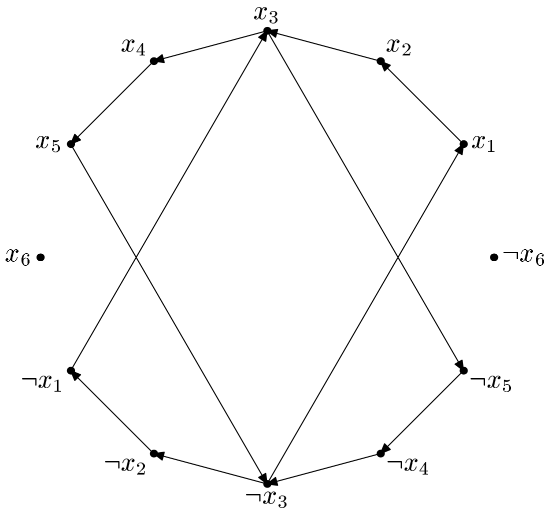

FIGURE 4. Le graphe associé à Φ_3

contenant x_3 présent dans ce graphe.

2.3. Les épines. On définit l'épine $S(F)$ d'une formule F en forme normale conjonctive comme l'ensemble des littéraux u tels qu'il existe une sous formule H satisfaisable de F for laquelle $H \wedge u$ n'est pas satisfaisable. Ainsi,

$$S(F) = \{u | \exists H \subset F, H \text{ est satisfaisable et } H \wedge u \text{ est insatisfaisable}\}.$$

Cet ensemble de littéraux d'une formule F correspond à un ensemble de sommets du graphe associé $G(F)$. Ainsi, on a le lemme suivant

Lemme 2.6. *Pour toute formule F sous forme $2CNF$,*

$$S(F) = \{u | u \overset{G(F)}{\rightsquigarrow} \neg u\}.$$

Ce lemme est montré dans l'appendice C.

3. A la recherche de cycles

Si les sommets représentant un littéral et sa négation sont dans un même cycle contradictoire, alors ils sont bien dans la même composante fortement connexe. C'est en exploitant cette idée que Chvátal et Reed [**10**] et Goerdt [**20**] ont obtenus les premiers résultats concernant le seuil de 2-SAT aléatoire.

Pour traiter le cas sous-critique dans lequel une formule contient moins de clauses que de variables, Chvátal et Reed ont introduit les *bicycles* dont ils montrent la présence dans le cas où il existe un cycle contradictoire dans le graphe associé à la formule. Il suffit alors de montrer que la probabilité de présence d'un bicycle est très faible dans le cas d'une formule sous-critique. Cette démonstration est présentée dans la section 2 du chapitre 3.

Dans le cas d'une formule sur-critique, Chvátal et Reed ont utilisé les *serpents*. Cette structure est une formule non satisfaisable qui contient donc un cycle contradictoire. Il reste alors à montrer qu'une formule sur-critique contient un serpent avec grande probabilité, ce qui est fait dans la section 3 du chapitre 3.

Goerdt [**20**] quant à lui, dans le cas sous-critique a considéré une forme dite *normalisée* des cycles contradictoires. Cette forme normalisée existe du fait de la dépendance de certains arcs. En effet, dans le graphe associé à une formule, il faut se souvenir que pour une clause, on introduit deux arcs, ceci permet de confirmer l'existence d'arcs utiles pour la *normalisation* des cycles contradictoires. Il reste alors à montrer que ces cycles sont presque sûrement absents d'un graphe associé à une formule sous-critique. Cette preuve est réalisée à l'aide d'arguments combinatoires.

Le cas sur-critique est traité de façon technique en montrant que l'espérance du nombre de cycles contenant une seule paire contradictoire (un littéral et sa négation) est grand et, grâce à la méthode du second moment, on montre qu'il y a un cycle présent dans chaque graphe.

Plutôt que de chercher les cycles contradictoires on peut chercher les composantes connexes contenant un littéral et sa négation.

4. Des utilisations d'arbres

Karp [**26**] a étudié les composantes fortement connexes dans le graphe aléatoire dirigé $D_{n,p}$ et a montré que si p est de la forme c/n avec c une constante, on voit apparaître un phénomène de seuil. En effet, pour $c < 1$, le graphe est composé de nombreuses composantes fortement connexe dont la taille est inférieure à $M \log n$, avec M une constante, tandis que si $c > 1$ le graphe contient une composante fortement connexe de taille asymptotiquement égale à Θn, où Θ est une constante qui ne dépend que de c.

Sachant que l'on peut transformer une formule 2-CNF en un graphe dirigé et que si dans ce graphe un littéral et sa négation appartiennent à la même composante fortement connexe alors la formule est insatisfaisable on va chercher à appliquer les résultats de Karp pour en déduire des bornes sur la satisfaisabilité. Malheureusement dans le graphe obtenu par la transformation les arêtes ne sont pas toutes deux à deux indépendantes contrairement à $D_{n,p}$ dans lequel c'est le cas. Aussi les résultats ne sont pas directement applicables mais des résultats similaires peuvent être obtenus.

Ainsi, Goerdt [**21**] a pu montrer que dans le graphe associé à une formule aléatoire les tailles des composantes fortement connexes sont petites et donc il est très peu probable que les sommets représentants un littéral et sa négation soit dans la même composante. La méthode utilisée est similaire à la méthode utilisée par Karp [**26**] puis reprise par Alon et Spencer [**3**] pour déterminer les composantes connexes dans le graphe aléatoire de Erdös-Rényi. Il s'agit d'analyser le comportement d'un algorithme qui développe sous la forme d'arbres l'ensemble des descendants d'un sommet. On a alors affaire à un processus de branchement dont l'analyse permet de conclure.

Dans le cas d'une formule 2-CNF aléatoire sur-critique, Verhoeven [**37**] utilise indirectement le fait que les composantes fortement connexes du graphe associé à la formule sont de tailles suffisantes pour que l'une d'entre elles contienne les sommets représentants un littéral et sa négation. Il montre en fait qu'avec grande probabilité il existe un littéral x tel que le sommet associé dans le graphe peut atteindre un grand nombre de sommets tandis que le sommet y représentant la négation du littéral est atteignable à partir d'un grand nombre de sommets. Alors, étant donné la taille de ces deux ensembles, il existe presque sûrement un arc entre ces deux ensembles permettant d'aller de x à y. Une légère variante de cette technique permet de montrer qu'il existe un cycle contradictoire dans le graphe. La preuve consiste à analyser un algorithme construisant un arbre de sommets atteignables à l'aide d'un processus de branchement particulier, de type *Galton-Watson*.

Toutefois, ce n'est pas en utilisant cette méthode que les résultats optimaux concernant le problème de 2-SAT aléatoire ont été obtenus mais grâce à une autre structure, l'*épine*.

5. Des arbres et des épines

Afin de trouver le comportement exact du seuil de transition en fonction de n du problème de 2-SAT aléatoire, Bollobás, Borgs, Chayes, Kim et Wilson [**7**] ont introduit l'*épine*. En contrôlant suffisamment la constitution de l'épine, on peut contrôler la satisfaisabilité de la formule, celle-ci étant insatisfaisable lorsqu'un littéral et sa négation appartiennent à l'épine (d'après le lemme 2.6). Ce contrôle de l'épine est effectué grâce à l'étude d'un graphe aléatoire non orienté avec une probabilité d'arête adéquate. Il s'agit tout d'abord de traduire l'épine en terme d'ensembles de sommets atteignables. Ensuite on transforme ces ensembles de sommets en arbres, pour cela il faut supprimer les cycles qui peuvent apparaître. Enfin, on montre que les tailles de ces arbres sont distribués comme les tailles des composante connexes du graphe aléatoire non orienté. Ces composantes sont bien connues, mais de nouveaux résultats concernant les graphes non orientés aléatoires sont nécessaires pour conclure. On obtient alors une borne inférieure du seuil de transition.

Afin d'obtenir une borne supérieure, Bollobás *et al.* étudient ce qu'ils nomment un *hourglass*. Cette structure est similaire à une réunion des graphes entrant et sortant d'un littéral, qui ont été utilisés par Goerdt [**21**] et Verhoeven [**37**]. Il est prouvé que ces structures sont liées à l'épine de la formule et donc que leur taille dépend la satisfaisabilité de la formule. L'analyse d'un algorithme permettant de développer ces hourglass nous donne les résultats nécessaires pour trouver les bornes supérieures correspondantes au seuil.

Un schéma de la preuve de ce résultat est présenté dans l'appendice C.

6. Un résultat en utilisant un autre graphe

Fernandez de la Vega [**12**] a obtenu un résultat sur la transition de phase en utilisant un graphe différent de celui introduit par Aspvall, Plass et Tarjan [**4**]. Le théorème est le suivant,

Théorème 2.7 (Fernandez de la Vega, 1992)**.** *Soit une formule 2-CNF aléatoire $F_{n,m}$. Si n et m tendent vers l'infini avec $m/n \geq c$, $c > 1$ constante fixée, alors, la probabilité $q(n)$ que la formule $F_{n,m}$ soit satisfaisable vérifie,*

$$q(n) \leq \exp\left(-f(c)n^{1/2}\right)$$

où $f(c)$ ne dépend que de c et est strictement positive pour $c > 1$.

Pour prouver ce théorème, Fernandez de la Vega transforme aussi sa formule en graphe mais cette fois le graphe n'est pas dirigé. Ce graphe contient $2n$ sommets, chacun correspondant à un littéral où la négation d'un littéral et les arêtes sont classées en deux sous-ensembles, E_1 et E_2. E_1 est l'ensemble des arêtes entre chaque littéral et sa négation tandis que E_2 correspond aux clauses de $F_{n,m}$ de manière naturelle. On va chercher les sommets l tels qu'il existe un cycle contenant l tel que les deux arêtes incidentes à l appartiennent à E_2 et les autres arêtes du cycle appartiennent alternativement à E_1 et E_2. Afin de montrer l'insatisfaisabilité d'une formule, il reste alors à montrer qu'il existe deux tels cycles, l'un correspondant à une certaine variable p tandis que l'autre correspond à $\neg p$. Ceci est fait via l'analyse d'un algorithme cherchant à construire de tels cycles. Cet algorithme résout une variante du problème consistant à savoir si un sommet est accessible dans un graphe aléatoire orienté. La variation étant due au lien particulier existant entre un sommet et sa négation. L'analyse est réalisée à l'aide de processus de branchement de type Galton-Watson.

CHAPITRE 3

Le seuil de 2-SAT aléatoire

1. Introduction

Dans ce chapitre nous allons étudier plus en détail le seuil que présente la probabilité de satisfaction des formules 2-CNF aléatoires qui a été évoqué dans le chapitre 1. Ainsi lorsqu'on étudie la probabilité de satisfaction des formules 2-CNF aléatoires (le modèle de probabilité sera défini plus précisément dans la suite) en fonction du nombre de clauses des formules, la courbe obtenue présente un phénomène de seuil. L'existence et la valeur de ce seuil ont été formellement déterminés, de façon indépendante, par Goerdt ([**20**]) et par Chvátal et Reed [**10**]. Le théorème prouvé est le suivant :

Théorème 3.1 (Chvátal et Reed, 1992, Goerdt 1992). *Soit* $F = F_{n,m}$ *une formule 2-CNF aléatoire, conjonction de* m *clauses différentes, sur* n *variables. Alors, quand* n *et* m *tendent vers l'infini,*

i. si $\frac{m}{n} \leq c$*, pour* $c < 1$ *fixé, alors* F *est presque sûrement satisfaisable ;*
ii. si $\frac{m}{n} \geq c$*, pour* $c > 1$ *fixé, alors* F *est presque sûrement insatisfaisable.*

On peut remarquer que grâce à ce qui a été dit dans la section 7 du chapitre 1, nous pouvons remplacer la formule $F_{n,m}$ du théorème par $F_{n,p}$, avec $p = c/(2n)$.

Dans ce chapitre nous allons présenter la démonstration de Chvátal et Reed.

2. Le cas des formules contenant peu de clauses

On commence par montrer la partie i du théorème 3.1. Il s'agit de montrer que les formules aléatoires contenant peu de clauses, relativement au nombre de variables sont presque sûrement satisfaisables.

Preuve de la partie i du théorème 3.1. Nous utilisons ici le modèle de formule aléatoire $F_{n,p}$, où chaque clause est présente avec probabilité $p = c/n$. Dans ce modèle, les formules obtenues contiennent $m = (1 + o(1))cn$ clauses (on rappelle que qu'ici une fonction $o(1)$ tend vers 0 quand le nombre de variables n d'une formule tend vers l'infini). Comme on a $c < 1$ fixé, on a bien $(1 + o(1))c < 1$.

Afin de montrer la partie i du théorème 3.1, il suffit de montrer les deux assertions suivantes,

a. avec une probabilité qui tend vers 0 quand n tend vers l'infini, une formule aléatoire contenant $(1 + o(1))cn$ clauses de taille deux sur n variables contient un bicycle ;
b. toute formule insatisfaisable dont toutes les clauses sont de taille deux contient un bicycle.

L'assertion b découle du lemme 2.3.

On montre a. Soit q la probabilité qu'une formule aléatoire avec m clauses contienne un bicycle. La probabilité qu'une formule aléatoire contienne un bicycle de taille r est inférieure à la probabilité que cette formule contienne l'équivalent d'un bicycle mais sans la condition de différence des variables. Alors on obtient,

$$\mathbb{E}[\text{Nombre de formules contenant un bicycle de taille } r] \leq n^r 2^r (2r)^2 p^{r+1}$$

car il y a n choix de variables, pour chaque variable on peut prendre la négation, il faut ensuite choisir les variables à chaque extrémité et enfin, on considère la probabilité de chaque clause. Alors, on a

$$\begin{aligned} q &\leq \sum_{s=2}^{n} n^s 2^s (2s)^2 m^{s+1} \left(\frac{1}{4\binom{n}{2}} \right)^{s+1} \\ &\leq \frac{2m}{n(n-1)} \sum_{s=2}^{n} s^2 \left(\frac{m}{n-1} \right)^s . \end{aligned}$$

Si $m = (1+o(1))cn$, alors $q = O(n^{-1})$, ce qui prouve a.

Ceci achève la démonstration de la partie i du théorème 3.1. □

3. Le cas des formules contenant plus de clauses que de variables

Preuve de la partie ii du théorème 3.1. Dans cette partie, les formules sont du type $F_{n,m}$. Soit un ensemble contenant n variables et t, dépendant de n tel que

$$\frac{t}{\ln n} \to \infty \tag{1}$$

mais par contre,

$$\frac{t}{n^{1/9}} \to 0. \tag{2}$$

On montre que, avec probabilité $1 - o(1)$, une formule aléatoire F contenant $(1+o(1))cn$ clauses de taille deux sur n variables avec $c > 1$, inclut toutes les clauses d'une certaine F_A. La preuve est une application de la méthode du second moment. On écrit $X = \sum X_A$, avec $X_A = 1$ si chaque clause de F_A apparaît exactement une fois dans F et $X_A = 0$ sinon. Nous montrons que

$$\mathbb{E}[X^2] \leq (1+o(1))\mathbb{E}[X]^2. \tag{3}$$

On considère deux serpents arbitraires A et B. On a $\mathbb{E}[X_A] = \mathbb{E}[X_B] = f(2t)$ (par définition un serpent est une suite de littéraux de taille $2t-1$ et la formule associée est de taille $2t$) avec

$$f(x) = \binom{m}{x} x! \left(\frac{1}{4\binom{n}{2}} \right)^x \left(1 - \frac{x}{4\binom{n}{2}} \right)^{m-x} = (1+o(1)) \left(\frac{m}{2n(n-1)} \right)^x$$

car $m = O(n)$ uniformément pour tout $x = O(n^\alpha)$ avec $\alpha < 1$. Si F_A et F_B partagent exactement i clauses, alors $\mathbb{E}[X_A X_B] = f(4t - i)$.

Pour un serpent A fixé et un serpent B choisi uniformément parmi tous les serpents, on note $p_i(n)$ la probabilité que les ensembles F_A et F_B aient exactement i clauses en commun. On peut montrer (voir [**10**], grâce à des arguments combinatoires et en considérant des graphes plutôt que des formules, que pour $1 \leq i \leq t-1$

$$p_i(n) < \frac{1300t^9}{n} \left(\frac{1}{2n} \right)^i , \tag{4}$$

et pour $1 \leq i \leq 2t$

$$p_i(n) < 18tn \left(\frac{1}{2n} \right)^i . \tag{5}$$

Supposons que F_A et F_B ont exactement i clauses en commun, alors, à l'aide de (1), (2) et en se souvenant que $\mathbb{E}[X_A X_B] = f(4t - i)$, on a

$$\mathbb{E}[X_A X_B] = (1+o(1))\mathbb{E}[X_A]\mathbb{E}[X_B] \left(\frac{2n(n-1)}{m} \right)^i$$

pour $0 \leq i \leq 2t$. Comme

$$\mathbb{E}[X^2] = \sum_A \sum_B \mathbb{E}[X_A X_B] \text{ et } \mathbb{E}[X]^2 = \sum_A \sum_B \mathbb{E}[X_A]\mathbb{E}[X_B]$$

il s'ensuit que

$$\mathbb{E}[X^2] = (1 + o(1))\mathbb{E}[X]^2 \sum_{i=0}^{2t} p_i(n) \left(\frac{2n(n-1)}{m}\right)^i.$$

Finalement, grâce à (4) puis (1),

$$\begin{aligned} \sum_{i=1}^{t-1} p_i(n) \left(\frac{2n(n-1)}{m}\right)^i &< \frac{1300t^9}{n} \sum_{i=1}^{t-1} \left(\frac{n-1}{m}\right)^i \\ &= o(1). \end{aligned}$$

Et enfin, (5) puis (2) nous permettent d'avoir,

$$\begin{aligned} \sum_{i=t}^{2t} p_i(n) \left(\frac{2n(n-1)}{m}\right)^i &< 18tn \sum_{i=t}^{2t} \left(\frac{n-1}{m}\right)^i \\ &= o(1). \end{aligned}$$

Et comme de plus on a $p_0(n) \leq 1$, ceci finit de montrer (3). Si alors on substitue $\mathbb{E}[X]$ à d dans l'inégalité de Chebyshev suivante,

$$\mathbb{P}\mathrm{r}[|X - \mathbb{E}[X]| \geq d] \leq \frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{d^2},$$

on conclut que X est non nulle avec grande probabilité, ce qui conclut la preuve du théorème. $\square$

CHAPITRE 4

Approximation de l'amplitude du seuil de 2-SAT, le cas sous-critique

1. Introduction

Dans le chapitre 3 nous avons montré qu'il existe une constante c telle que, quel que soit ε fixé, si une formule 2-CNF aléatoire contient moins de $(c-\varepsilon)n$ clauses alors elle est presque sûrement satisfaisable, tandis que si elle contient plus de $(c+\varepsilon)n$ clauses, alors elle est presque sûrement insatisfaisable. De plus, nous avons vu que $c = 1$. Dans ce chapitre et le suivant nous allons raffiner ce résultat et approximer la largeur maximale de l'intervalle $I(n)$ tel que pour $m \in I(n)$, les probabilités que $F_{n,m}$ soit satisfaisable et que $F_{n,m}$ soit insatisfaisable soient toutes deux strictement différentes de 0. Dans un premier temps nous allons exposer les résultats obtenus dans le cas de formules sous-critiques, c'est à dire telles que $c_2 < 1$. Puis dans le chapitre suivant nous étudierons le cas de formules sur-critique, où $c_2 > 1$.

Nous commençons par montrer le théorème suivant dû à Goerdt [**21**] et concerne le comportement du graphe en dessous du seuil.

Théorème 4.1 (Goerdt, 1999). *Soit $F = F_{n,p}$ une formule 2-CNF aléatoire dans laquelle chaque clause de l'ensemble des $4\binom{n}{2}$ clauses possibles sur n variables est présente avec la probabilité $p \leq (1-(\ln n)^{-1/2})/2n$. Alors $F_{n,p}$ est presque sûrement satisfaisable.*

Ce théorème utilise le modèle de graphe présenté dans la section 5 du chapitre 1 et expose un algorithme qui permet de vérifier l'absence de cycle contradictoire. La démonstration de ce théorème fait l'objet de la section 2.

Et enfin, nous verrons dans la section 3 qu'une très bonne approximation de la valeur inférieure de l'intervalle de seuil de 2-SAT aléatoire peut-être obtenue en effectuant un calcul plus précis dans la démonstration de la partie i du théorème 3.1. On obtient alors le théorème suivant,

Théorème 4.2. *Soit $F = F_{n,p}$ une formule 2-CNF aléatoire dans laquelle chaque clause de l'ensemble des $4\binom{n}{2}$ clauses possibles sur n variables est présente avec la probabilité $p \leq (1-w(n)n^{-1/3})/2n$, $w(n)$ fonction croissante non bornée. Alors $F_{n,p}$ est presque sûrement satisfaisable.*

Nous pourrons remarquer que les démonstrations de Goerdt et la notre utilisent une idée commune. En effet, toutes deux cherchent à déterminer la probabilité d'existence d'un cycle contradictoire. Pour cela, dans les deux cas, il s'agit de déterminer, pour un littéral u fixé, la probabilité qu'un chemin existe entre u et $\neg u$ dans le graphe associé à la formule. Pour cela, nous allons exposer des algorithmes construisant l'arbre contenant des sommets accessibles depuis u et l'arbre des sommets à partir desquels $\neg u$ est accessible. Si ces arbres sont suffisamment petits, alors ils sont presque sûrement disjoints dans le graphe aléatoire associé à la formule. Par contre à partir d'une certaine taille, il existe presque sûrement un chemin menant d'un arbre à l'autre et donc de u vers $\neg u$.

Dans la suite, l'ensemble des sommets accessibles à partir d'un sommet u sera noté $D^{+}(u)$. Cet ensemble est appelé *arbre sortant* de u (bien que ce ne soit un arbre car il peut contenir des cycles). On considérera aussi l'ensemble des sommets à partir desquels on peut atteindre u et on le note $D^{-}(\neg u)$. Ce second ensemble est l'*arbre entrant* de u.

On peut remarquer que, par construction du graphe, $D^{-}(\neg u)$ est constitué des négations de tous les littéraux qui constituent $D^{+}(u)$ (on se souviendra que l'existence de l'arête $v \to w$ implique l'existence de l'arête $\neg w \to \neg v$).

2. Plus qu'une remarque concernant 2-SAT aléatoire

Nous allons en fait montrer ici une version un petit peu plus forte du théorème 4.1 dont la démonstration est due à Fernandez de la Vega [**11**]. Cette démonstration est essentiellement la même que celle présentée par Goerdt [**21**]. Le théorème est le suivant,

Théorème 4.3. *Soit $F = F_{n,p}$ une formule 2-CNF aléatoire dans laquelle chaque clause de l'ensemble des $4\binom{n}{2}$ clauses possibles sur n variables est présente avec la probabilité $p \leq (1 - n^{-1/8}\ln n)/2n$. Alors $F_{n,p}$ est presque sûrement satisfaisable.*

Afin de montrer le théorème 4.3, nous allons présenter l'algorithme *fanning out procedure*. Cet algorithme travaille sur le graphe associé à une formule tel qu'il a été défini dans la section 5 du chapitre 1. Nous supposons l'ordre suivant sur les littéraux : $\{u_1, \neg u_1\} < \{u_2, \neg u_2\} < \ldots < \{u_n, \neg u_n\}$.

2.1. L'algorithme *fanning out procedure*. L'algorithme prend une formule F sous forme 2-CNF en entrée et un littéral u de la formule F et produit un sous-ensemble $X(u)$ qui est sous-ensemble de l'ensemble $D^{+}(u)$ de tous les littéraux accessibles à partir de u dans le graphe de la formule F. $D^{+}(u)$ est aussi appelé l'*arbre sortant de u*. L'algorithme utilisé nous assure que $X(u)$ ne contient pas un littéral v et sa négation. De plus, cet algorithme construit une suite de couples $(A_0, B_0), (A_1, B_1), \ldots, (A_r, B_r)$ où A_i contient les littéraux accessibles à partir de u dans le graphe et dont l'algorithme a déjà atteint les successeurs, tandis que B_i contient tous les littéraux accessibles depuis u et déjà atteints par l'algorithme. On a donc $A_i \subset B_i$ pour tout $i < r$, et $A_r = B_r = X(u)$.

Fanning out procedure

1. $A_0 = \emptyset$, $B_0 = \{u\}$ et $i := 0$;
2. Tant que $A_i \neq B_i$:
   a. $i := i + 1$;
   b. Soit w le plus petit littéral de $B_i \backslash A_i$;
   c. $A_{i+1} := A_i \cup \{w\}$;
   d. B_{i+1} est égal à B_i augmenté de tous les littéraux v vérifiant :
      i. le graphe associé à F a une arête $w \to v$ et *pas* d'arête $w \to \neg v$. Aucun des littéraux v et $\neg v$ n'appartient à B_i ;
      ii. le graphe associé à F a une arête $w \to v$ et une arête $w \to \neg v$, x est la variable associée à v. Aucun des littéraux x et $\neg x$ n'appartient à B_i ;

Remarque 2.4. Il est clair que $X(u)$ est un sous ensemble de $D^{+}(u)$ car par construction $X(u)$ ne peut contenir un littéral et son complémentaire.
On définit $Y(\neg u)$ comme l'ensemble contenant l'arête $v \to w$ si et seulement si $X(u)$ contient l'arête $\neg w \to \neg v$. $Y(\neg u)$ est un sous-ensemble de l'ensemble $D^{-}(\neg u)$. La

construction de $X(u)$ et de $X(\neg u)$ nous permet d'obtenir aussi les ensembles $Y(u)$ et $Y(\neg u)$. On remarque que $Y(\neg u)$ contient les négations des littéraux présents dans $X(u)$, ces deux ensembles sont donc disjoints (en effet l'algorithme garantit que si pour un littéral v on a $v \in X(u)$, alors $\neg v \notin X(u)$).

2.2. Analyse du fanning out process. Cet algorithme nous permet de construire, pour tout littéral u apparaissant dans une formule un ensemble de sommets accessibles à partir de u dans le graphe associé à la formule et aussi, toujours dans ce graphe, un ensemble de sommets à partir desquels on peut atteindre $\neg u$. Il nous faut maintenant estimer la taille de ces ensembles. Pour cela nous démontrons la proposition suivante,

Proposition 4.5. *On pose* $p \leq (1-n^{-1/8}\ln n)/2n$*. Soit une formule* $F = F_{n,p}$ *alors on a*

$$\mathbb{P}\mathrm{r}\left[F : \exists v \textit{ tel que } |X(v)| \geq \frac{16n^{\frac{1}{4}}}{\ln n}\right] \leq o\left(\frac{1}{n}\right). \tag{6}$$

Preuve de la proposition 4.5. Analysons le comportement de l'algorithme *fanning out procedure* en prenant comme entrée une formule aléatoire $F = F_{n,p}$. Étant donné que pour $i < r$ on a $A_i \subset B_i$ $(A_i \neq B_i)$, w est le plus petit littéral de $B_i \backslash A_i$ et u est un littéral tel que $u \notin B_i$ et $\neg u \notin B_i$, on calcule la probabilité que u soit dans B_{i+1}. C'est exactement la probabilité que $\neg w \vee u$ ou $\neg w \vee \neg u$ soient des clauses de F, c'est à dire $q = 2p - p^2$. Il y a $n - |B_i|$ variables x telles que x ou $\neg x$ (mais pas les deux) peuvent être ajoutées à B_i. Ainsi, la distribution du nombre de littéraux ajoutés à B_i, connaissant $|B_i|$ suit une loi binomiale $Bin(n - |B_i|, q)$ de paramètres $n - |B_i|$ et q.
On définit ensuite une suite de variables aléatoires $(Z_i)_{i\in\mathbb{N}}$ comme suit :

$Z_0 = 1$
$Z_{i+1} = Z_i + Bin(n - Z_i, q)$ pour $i > 0$.

Il est aisé de vérifier que, tant que $B_i \geq i$, Z_i et B_i ont la même distribution. Soit v un littéral de F. On peut trouver la distribution de $|X(v)|$,

$$\begin{aligned}\mathbb{P}\mathrm{r}[|X(v)| = t] &= \mathbb{P}\mathrm{r}[|B_0| > 0 \text{ et } |B_1| > 1 \text{ et } \ldots \text{ et } |B_{t-1}| > t-1 \text{ et } |B_t| = t] \\ &= \mathbb{P}\mathrm{r}[\min\{i | Z_i = i\} = t) \\ &\leq \mathbb{P}\mathrm{r}[Z_t = t].\end{aligned}$$

Si maintenant on définit une autre suite de variables aléatoires $(W_i)_{i\in\mathbb{N}}$ comme suit :

$W_0 = 1$
$W_{i+1} = W_i + Bin(n, q)$ pour $i > 0$,

on peut remarquer que $\mathbb{P}\mathrm{r}[Z_t \geq t] \leq \mathbb{P}\mathrm{r}[W_t \geq t]$. Et, de plus, pour $t \geq 1$, la distribution de W_t est une loi binomiale $Bin(tn, q)$ de paramètres tn et q. Si on pose $p = (1-\varepsilon)/2n$, en utilisant une borne de Hoeffding-Chernoff, on a

$$\mathbb{P}\mathrm{r}[W_t \geq t] \leq \exp\left(-\frac{\varepsilon^2 t}{3}\right).$$

On choisit maintenant $\varepsilon = n^{-1/8}\ln n$ et $t = 16n^{1/4}/\ln n$ comme l'a fait Fernandez de la Vega [**11**] et on obtient,

$$\begin{aligned}\mathbb{P}\mathrm{r}[W_t \geq t] &\leq \exp\left(-\frac{16\ln n}{3}\right) \\ &= o\left(\frac{1}{n^2}\right).\end{aligned}$$

Ce qui achève la démonstration de (6). □

Remarque 2.6. Du fait de la construction particulière du graphe, le résultat du lemme 4.5 nous permet aussi d'avoir

$$\mathbb{P}\mathrm{r}\left[F : \exists v \text{ tel que } |Y(\neg v)| \geq \frac{4n^{1/4}}{\ln n}\right] \leq o\left(\frac{1}{n}\right).$$

Il faut maintenant faire le lien entre les ensembles que l'algorithme construit et la satisfaisabilité des formules. Nous utilisons pour cela la proposition suivante, en nous souvenant qu'une formule sous forme 2-CNF est insatisfaisable si et seulement si son graphe associé contient un cycle contradictoire.

Proposition 4.7. *Le graphe associé à la formule F contient un cycle contradictoire contenant u et $\neg u$ si et seulement si il existe une arête de $X(u)$ vers $Y(\neg u)$ et une arête de $X(\neg u)$ vers $Y(u)$ dans le graphe.*

Preuve de la proposition 4.7. Si il existe un littéral u tel que le graphe associé à F contienne une arête reliant un sommet de $X(u)$ à un sommet de $Y(\neg u)$ et une autre arête reliant un sommet de $X(\neg u)$ à un sommet de $Y(u)$, alors u et $\neg u$ font partie de la même composante fortement connexe du graphe et le théorème 1.5 permet de conclure à l'insatisfaisabilité de la formule F. La réciproque n'est pas immédiate car $X(u)$ n'est qu'un sous-ensemble de $D^+(u)$. Supposons que la formule soit insatisfaisable et donc que le graphe associé contienne un cycle $(u, v_1, \ldots, v_r, \neg u, w_1, \ldots, w_s, u)$ dans lequel tous les littéraux sauf u apparaissent une seule fois. Soit v_i le premier littéral du cycle qui n'appartient pas à $X(u)$. Ce littéral est accessible à partir de u car il appartient au cycle, c'est donc que $\neg v_i$ appartient à $X(u)$, donc par construction v_i appartient à $Y(\neg u)$. Il existe donc une arête entre $X(u)$ et $Y(\neg u)$. De même on montre que dans ce cas il existe une arête entre $X(\neg u)$ et $Y(u)$, ce qui achève la démonstration de la proposition. □

Grâce à cette dernière proposition, si on montre que pour tout littéral u de la formule $F = F_{n,p}$, avec $p \leq (1 - n^{-1/8} \ln n)/2n$, il n'existe pas d'arête entre $X(u)$ et $Y(\neg u)$, ni entre $X(\neg u)$ et $Y(u)$, alors on aura montré le théorème 4.3.

Preuve du théorème 4.3. Soit $F = F_{n,p}$ une formule sous forme 2-CNF, avec $p \leq (1 - n^{-1/8} \ln n)/2n$. D'après la proposition 4.5, F est, avec grande probabilité, telle qu'on ait, pour tout littéral v,

$$|X(v)| \leq \frac{16n^{1/4}}{\ln n} \text{ et } |X(\neg v)| \leq \frac{16n^{1/4}}{\ln n}. \tag{7}$$

Soit donc une formule F telle que le graphe associé vérifie les conditions 7 pour tout littéral. Soit u un littéral de cette formule. Par définition, les ensembles $X(u)$ et $Y(\neg u)$ sont disjoints. Pour avoir un cycle contradictoire, il nous faut un arc de $X(u)$ vers $Y(\neg u)$ et un autre de $X(\neg u)$ vers $Y(u)$, c'est à dire une clause dont les deux littéraux appartiennent à $Y(\neg u)$ et une autre dont les deux littéraux appartiennent à $Y(u)$.

Pour un littéral u de F, on cherche à connaître la probabilité qu'une clause contienne deux littéraux de $Y(\neg u)$, sachant que $|Y(\neg u)| = |X(u)| \leq \frac{4n^{1/4}}{\ln n}$ et une probabilité similaire pour $Y(u)$. On différencie alors deux cas,

i. si $X(v) \cap X(\neg v) = \emptyset$, alors pour obtenir les arcs, il faut avoir deux clauses indépendantes, chacune d'entre elles étant tirée uniformément parmi $O\left(\frac{n^{1/2}}{\ln^2 n}\right)$ clauses. Ceci arrive avec probabilité $O\left(\frac{1}{n \ln^4 n}\right)$.
ii. Si $X(v) \cap X(\neg v) \neq \emptyset$, alors un cycle contradictoire est présent si on a au moins une clause choisie parmi $O\left(\frac{n^{1/2}}{\ln^2 n}\right)$ clauses. Or $X(u)$ et $X(\neg u)$ ne sont pas disjoints avec probabilité $O\left(\frac{1}{n^{1/2} \ln^2 n}\right)$.

Ainsi, la probabilité d'existence d'un cycle contradictoire contenant les littéraux u et $\neg u$ est de l'ordre de $o(1/n)$ et la probabilité que la formule F soit insatisfaisable est de l'ordre de $o(1)$.

Soit $F = F_{n,p}$ une formule aléatoire. Nous venons alors de montrer que la probabilité $\mathbb{P}\mathrm{r}\left[F \text{ est satisfaisable } | F : \forall u, |X(u)| \leq \frac{16n^{1/4}}{\ln n}\right]$ est de l'ordre de $1-o(1)$. La proposition 4.5 indique que la probabilité $\mathbb{P}\mathrm{r}\left[F : \forall v \text{ tel que } |X(v)| \leq \frac{4n^{1/4}}{\ln n}\right]$ est de l'ordre de $1-o(1)$. On en déduit alors

$$\begin{aligned}
\mathbb{P}\mathrm{r}[F \text{ est satisfaisable}] &\geq \mathbb{P}\mathrm{r}\left[F \text{ est satisfaisable et } F : \forall u, |X(u)| \leq \frac{16n^{1/4}}{\ln n}\right] \\
&\geq \mathbb{P}\mathrm{r}\left[F \text{ est satisfaisable } | F : \forall u, |X(u)| \leq \frac{16n^{1/4}}{\ln n}\right] \times \\
&\quad\ \mathbb{P}\mathrm{r}\left[F : \forall v \text{ tel que } |X(v)| \leq \frac{4n^{1/4}}{\ln n}\right] \\
&\geq 1 - o(1).
\end{aligned}$$

Ceci achève la démonstration du théorème 4.3. □

Nous pouvons alors nous intéresser à ce qui se passe de l'autre côté du seuil, lorsque la formule est insatisfaisable avec grande probabilité. Toutefois, auparavant, nous proposons un raffinement du raisonnement de Chvátal et Reed [**10**] (voir aussi le chapitre 3) qui permet l'obtention d'une borne inférieure pour le seuil de satisfaisabilité.

3. Une amélioration à partir des bicycles

Preuve du théorème 4.2. On se souvient que dans la section 2 du chapitre 3, on souhaitait borner la probabilité qu'une formule contenant m clauses contienne un bicycle. On avait :

$$q \leq \frac{2m}{n(n-1)} \sum_{s=2}^{n} s^2 \left(\frac{m}{n-1}\right)^s.$$

Or le choix qui a été fait en prenant $m = (1+o(1))cn$, avec $c < 1$ permet d'obtenir $q = O(n^{-1})$ tandis que $q = o(1)$ suffit. Un long et fastidieux calcul (qui est effectué dans l'appendice B) nous donne

$$q \leq \frac{2}{n} \frac{c^2}{(1-c)^3} \left[4c - 3c^2 + c^3 - (n+1)^2 c^n + (2n^2 + 2n - 1)c^{n+1} - n^2 c^{n+2}\right].$$

On souhaite trouver c tel que $q = O(1/w(n))$ avec $w(n)$ tendant vers l'infini arbitrairement lentement quand n tend vers l'infini. On sait que $c = (1 + o(1))$, il nous faut donc préciser la valeur du terme $o(1)$. Après calcul, on remarque qu'en choisissant $c = 1 - \left(\frac{w(n)}{n}\right)^{1/3}$ on obtient en effet le résultat annoncé dans le théorème 4.2. □

CHAPITRE 5

Approximation de l'amplitude du seuil de 2-SAT, le cas sur-critique

1. Introduction

Concernant l'approximation de la probabilité de satisfaisabilité d'une formule 2-SAT aléatoire au dessus du seuil, plusieurs résultats ont été obtenus concernant ce problème. Le premier a été obtenu par Fernandez de la Vega [**12**], où il énonce le théorème 2.7. On ne montrera pas ce théorème ici, mais il est intéressant car comme cela a été dit dans le chapitre 2 c'est le seul résultat concernant 2-SAT utilisant un graphe différent de celui présenté dans la section 5 du chapitre 1.

Par ailleurs, un autre résultat a été obtenu par Verhoeven [**37**] en présentant un algorithme qui explore le graphe classique à la recherche de cycles contradictoires. Cette méthode permet de montrer le théorème suivant,

Théorème 5.1. *Soit $F = F_{n,p}$ une formule 2-CNF aléatoire dans laquelle chaque clause de l'ensemble des $4\binom{n}{2}$ clauses possibles sur n variables est présente avec la probabilité $p \geq (1+n^{-1/4})/2n$. Alors $F_{n,p}$ est presque sûrement insatisfaisable.*

Ce théorème est similaire au théorème 4.3 mais concerne les formules dont le nombre de clauses est supérieur au nombre de variables, c'est à dire celles qui sont presque sûrement insatisfaisables d'après Goerdt [**20**], Chvátal et Reed [**10**] et Fernandez de la Vega [**12**]. Ces derniers ont montré que pour une formule 2-CNF aléatoire F sur n variables, pour tout $\varepsilon > 0$ fixé, si F contient plus de $(1+\varepsilon)n$ clauses, alors F est presque sûrement insatisfaisable (voir le théorème 3.1). Le résultat donné ici permet de préciser cette propriété en exprimant ε comme une fonction de n.

Une preuve de ce théorème consiste à exhiber un algorithme tel que, pour une formule $F = F_{n,p}$ tirée aléatoirement avec une probabilité $p \geq (1+n^{-1/4})/2n$, celui-ci trouve un cycle contradictoire avec grande probabilité dans le graphe associé à F. Dans ce cas, la formule est insatisfaisable (voir la section 5 du chapitre 1) et le théorème est alors prouvé.

Nous présentons un tel algorithme dans la section 2, puis son analyse est faite dans la section 3.

2. Notations et algorithme

2.1. Notations. Comme précédemment, $F = F_{n,p}$ est une formule 2-CNF aléatoire et on notera $G(F)$ le graphe associé tel qu'il a été défini dans la section 5 du chapitre 1. Dans la suite, $w(n)$ représentera une fonction quelconque, tendant vers l'infini quand n tend vers l'infini, $\varepsilon = \varepsilon(n)$ est une fonction qui sera définie ultérieurement. Enfin, soit $\delta = \varepsilon/4(1+\varepsilon)$ et on pose $N = 2n(1-\delta)$ et $1+\theta = (1+\varepsilon)(1-\delta)$. On a alors $\theta = 3\varepsilon/4$. Ce choix de δ sera justifié dans la suite.

2.2. L'algorithme d'accessibilité. L'algorithme présenté maintenant permet de rechercher des cycles contradictoires dans le graphe $G(F)$. Afin d'atteindre ce but, on va construire deux ensembles aléatoires de sommets du graphe tout en vérifiant si dans graphe aléatoire il existe des arcs les reliant. Pour cela on choisit arbitrairement une variable u de la formule et on tente de déterminer s'il existe un chemin dans $G(F)$ d'une part entre le sommet u et le sommet $\neg u$ et d'autre part entre $\neg u$ et u. Pour déterminer s'il existe un chemin entre u et $\neg u$ dans $G(F)$, on va considérer un sous-ensemble de l'ensemble de sommets accessibles à partir de u et un sous-ensemble de l'ensemble des sommets à partir desquels $\neg u$ est accessible. L'algorithme présenté ici développe ces sous-ensembles sous la forme d'arbres. Si l'arbre issu de u et l'arbre issu de $\neg u$ ont des frontières (les sommets appartenant à ces frontières n'ont pas encore été explorés) de taille suffisante, le graphe étant probabiliste, il existe presque sûrement un chemin de u vers $\neg u$. Il n'est pas nécéssaire de développer ces arbres entièrement mais de telle manières que leurs frontières soient de taille suffisante. Il faut toutefois remarquer que le graphe $G(F)$ est tel que pour chaque clause de F on a introduit deux arêtes, ainsi d'un arbre des sommets accessibles à partir de u on peut aussitôt déduire l'arbre des sommets à partir desquels on peut atteindre $\neg u$. Il suffit alors que l'algorithme développe un arbre de sommets accessibles à partir de u et un arbre de sommets accessibles à partir de $\neg u$.

Cet algorithme procède comme suit. Soit u un sommet aléatoire de $G(F)$. Soit $B_1^+(u)$ l'ensemble constitué des sommets de $G(F)$ qui sont accessibles à partir de u dans $G(F)$, soit $B_1^+(\neg u)$ l'ensemble des sommets de $G(F)$ qui sont accessibles à partir de $\neg u$ dans $G(F)$. Si $B_1^+(u)$ et $B_1^+(\neg u)$ sont de taille suffisante, on arrête l'algorithme et on vérifie s'il existe un arc entre ces deux ensembles dans le graphe aléatoire. Ces deux ensembles sont aléatoires et leurs voisins n'ont pas encore été explorés, aussi les arcs existent avec grande probabilité et il existe un cycle contradictoire contenant u et $\neg u$, la formule est insatisfaisable. Si $B_1^+(u)$ et $B_1^+(\neg u)$ ne sont pas de taille suffisante, alors on construit les ensembles $B_2^+(u)$ et $B_2^+(\neg u)$. $B_2^+(u)$ est égal à $B_1^+(u)$ augmenté des sommets accessibles à partir d'un sommet choisi aléatoirement dans $B_1^+(u)$ (ces sommets sont pris dans $G(F)$ privé de $B_1^+(u)$) et privé de ce sommet. De la même façon, $B_2^+(\neg u)$ est constitué des sommets de $B_1^+(\neg u)$ augmenté de l'ensemble des sommets accessibles à partir de l'un des sommets de $B_1^+(\neg u)$ choisi aléatoirement et privé de ce sommet. On continue de la même manière tant que les ensembles n'ont pas atteint la taille voulue. Si pour $i > 0$, on a construit $B_i^+(u)$ et $B_i^+(\neg u)$ mais que l'on ne peut pas construire $_{i+1}^+(u)$ ou $B_{i+1}^+(\neg u)$ alors l'algorithme s'arrête et on recommence avec un nouveau littéral v qui n'a pas été traité. Soit j le premier entier tel que $B_j^+(u)$ et $B_j^+(\neg u)$ soient de taille suffisante, on pose alors $X(u) = B_j^+(u)$ et $X(\neg u) = B_j^+(\neg u)$.

Il faut remarquer que si nous procédons exactement de cette façon, à chaque étape de l'algorithme le nombre de sommets parmi lesquels on cherche des successeurs va diminuer. Ce comportement est illustré par les figures 5 et 6 dans lesquelles on remarque la diminution du nombre de sommets libres. Afin d'éviter ce type de comportement, nous allons utiliser une technique déjà utilisée par Ajtai, Komlós et Szemerédi [**2**]. Il s'agit au début de l'algorithme de réserver un certain nombre de sommets parmi lesquels on ne regardera pas pour trouver des successeurs (cet ensemble sera noté H). Par contre, à la fin de chaque étape, on remplacera les sommets choisis comme successeurs lors de l'étape par le même nombre de sommets de H. Cette stratégie permet de chercher des successeurs parmi un ensemble de sommets de cardinal constant et est illustrée par les figures 7, 8 et 8. Comme nous le verrons par la suite, cette stratégie ne nous pénalise pas trop et avec grande probabilité on n'épuise pas l'ensemble H avant d'avoir terminé l'algorithme. Nous présentons maintenant l'algorithme de manière plus formelle. Il est décomposé en

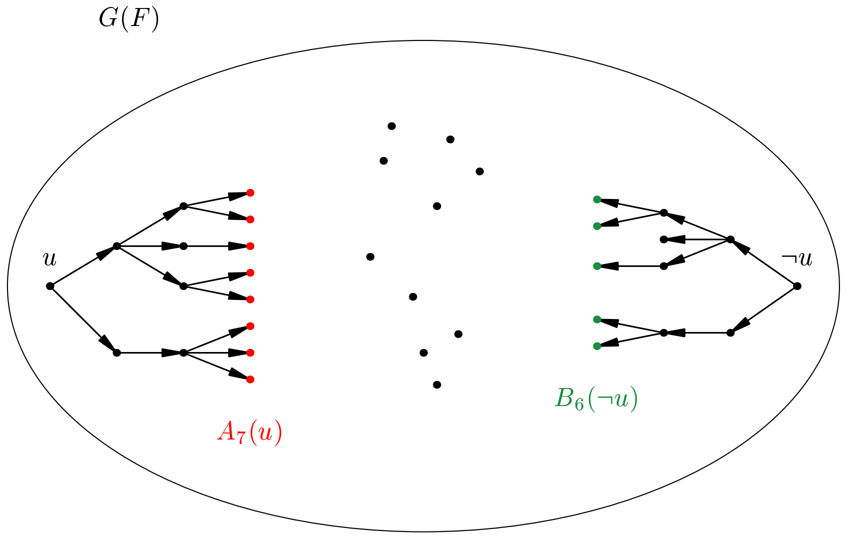

FIGURE 5. $G(F)$ avec $B_7^+(u)$ et $B_6^+(\neg u)$ développés.

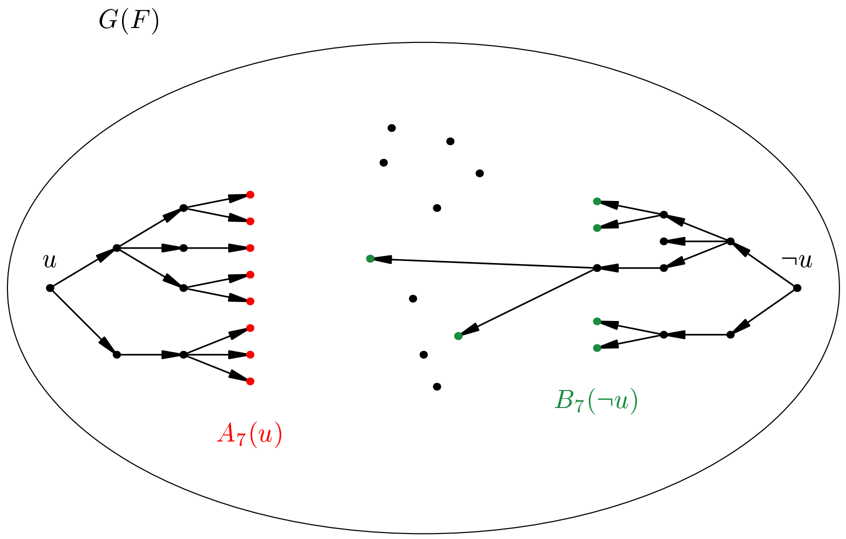

FIGURE 6. $G(F)$ avec $B_7^+(u)$ et $B_7^+(\neg u)$ développés.

deux parties **MAIN** et **TREE**, la première choisit un littéral et teste les ensembles générés par la seconde qui développe les arbres.

Algorithme MAIN

1. Soit H un sous-ensemble aléatoire de $V(F)$, contenant $2n-N$ sommets.
2. $T := V(F) - H$, $INF := faux$
3. **Tant que** $INF = faux$
   a. Soit u un sommet aléatoire de T tel que $\neg u \in T$; on pose $B_0^+(u) := \{u\}$;
   b. On exécute TREE pour $u \to X(u),\ X(\neg u)$;
4. On teste pour savoir s'il existe un chemin entre $X(u)$ et $Y(\neg u)$;

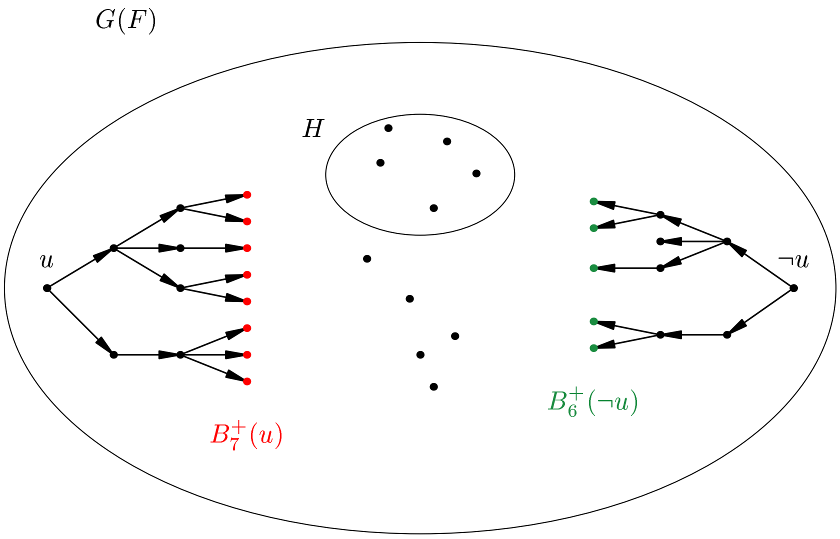

FIGURE 7. Le graphe avec $B_7^+(u)$ et $B_6^+(\neg u)$

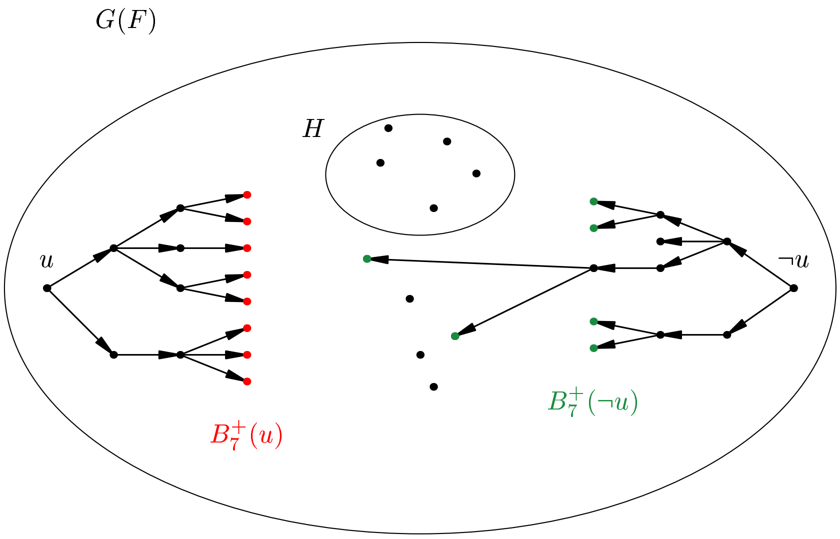

FIGURE 8. Construction de l'ensemble $B_7^+(\neg u)$

5. On teste pour savoir s'il existe un chemin entre $X(\neg u)$ et $Y(u)$;

Algorithme TREE

1. $END := faux$; $i := 0$; $B_0^+(u) := \{u\}$; $B_0^+(\neg u) := \{\neg u\}$;
2. **Tant que** $END = faux$,
    - Soit $m_1 := |B_i^+(u)|$ et $m_2 := |B_i^+(\neg u)|$. Alors $B_i^+(u) := \{x_1, x_2, \ldots, x_{m_1}\}$, $B_{i+1}^+(u) := \emptyset$ et $B_i^+(\neg u) := \{y_1, y_2, \ldots, y_{m_2}\}$, $B_{i+1}^+(\neg u) := \emptyset$
    - **si** $(|B_i^+(u)| \geq w(n)\sqrt{2n}$ **et** $|B_i^+(\neg u)| \geq w(n)\sqrt{2n})$ **alors** $END := true$ **et** $INF := true$
    - **si** $(m_1 = 0$ **ou** $m_2 = 0)$ **alors** $END := true$
    - **alors**

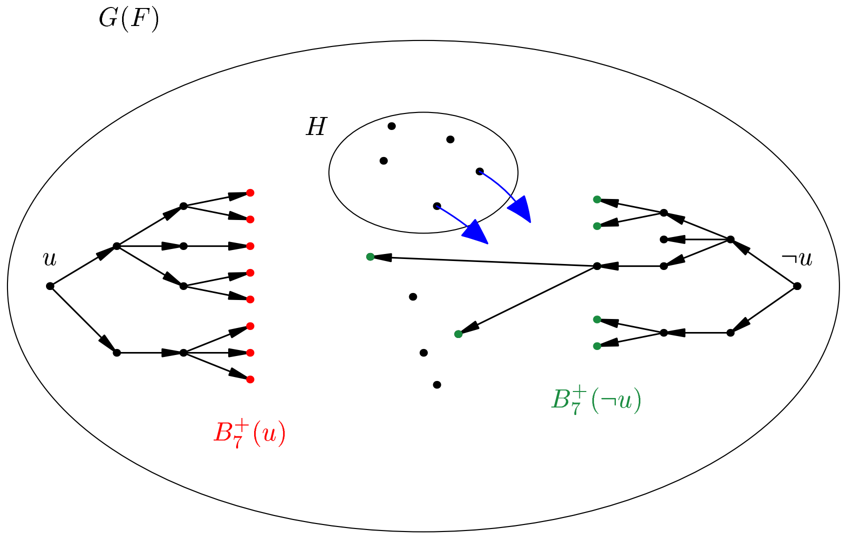

FIGURE 9. Rééquilibrage du nombre de sommets libres

- $B^+_{i+1}(u) := B^+_{i+1}(u) \cup (\Gamma^+(x_1)) \cap T$
- $B^+_{i+1}(\neg u) := B^+_{i+1}(\neg u) \cup (\Gamma^+(y_1)) \cap T$
- On remplace les ensembles de sommets $\Gamma^+(x_1)$ et $\Gamma^+(y_1)$ dans T par $|\Gamma^+(x_1)| + |\Gamma^+(y_1)|$ sommets de H.

Comme il a été noté précédemment les algorithmes travaillent sur l'ensemble des sommets privés d'un sous-ensemble, ainsi la distribution du nombre d'arêtes de chaque sommet reste inchangée au cours de l'exécution de l'algorithme et l'analyse peut se faire au moyen de processus de *Galton-Watson*. Une introduction à ces processus ainsi qu'une présentation des résultats qui seront utiles ici peuvent être trouvés dans l'appendice A.

On remarque qu'en utilisant les notations qui ont été introduites précédemment, c'est à dire en notant N le nombre de sommets sur lequel on travaille, la probabilité d'arête du graphe sur N sommets est $(1+\theta)/N$. Ceci sera utilisé dans la suite.

3. Analyse de l'algorithme

3.1. Analyse d'un processus de branchement. Nous considérons ici un processus de Galton-Watson dont la distribution de reproduction est la distribution binomiale $\mathrm{Bin}(N, (1+\theta)/N)$ avec N et θ fixés. De plus, Z_n et S représenteront respectivement la taille de la $n^{\text{ième}}$ génération et la taille totale du processus ($S = \sum_{n\geq 0} Z_n$). S peut éventuellement être infini.

En plus de cette représentation classique des processus de branchement, nous allons aussi considérer la suite d'arbres $(T_i)_{i\geq 0}$ définie comme suit : T_0 est un arbre constitué d'un seul sommet et, pour tout $i \geq 0$, T_{i+1} est obtenu à partir de T_i en développant un sommet qui n'a pas encore été exploré (lorsque tous les sommets ont été développés, T_{i+1} est vide). L'ensemble des sommets non encore explorés de T_i sera appelé la *frontière* de T_i et sera noté B_i. L'ordre dans lequel on développe les sommets n'a pas d'importance par rapport à l'analyse que nous effectuons, aussi nous les développons dans l'ordre dans lequel ils sont produits. Cela permet de créer une cohérence entre les deux représentations : pour chaque $i \geq 0$, Z_{i+1} est le nombre de sommets de profondeur $i+1$ dans l'arbre lorsque tous les sommets de profondeur i ont été développés.

Nous nous intéressons maintenant au développement du processus et à la probabilité d'extinction dans notre cas particulier. Nous énonçons donc le lemme suivant,

Lemme 5.2. *Soit un processus de Galton-Watson GW dont la loi de reproduction est la distribution binomiale $Bin(N,(1+\theta)/N)$. Alors,*

$$\mathbb{P}\mathrm{r}\,[GW \textit{ ne s'éteint pas}] \geq \frac{\theta}{(1+\theta)^2}.$$

Conditionnellement à l'extinction du processus, on a

$$\mathbb{E}[|S|] \leq 1+\theta.$$

Preuve du lemme 5.2. Soit q la probabilité d'extinction d'un processus de type Galton-Watson de loi de reproduction la distribution binomiale $\mathrm{Bin}(N,(1+\theta)/N)$. On remarque que la moyenne de la distribution de reproduction est supérieure à 1, donc le processus a une probabilité non nulle de survie. Si nous notons f la fonction génératrice de la loi de probabilité $\mathrm{Bin}(N,(1+\theta)/N)$, alors q est la solution dans $]0,1[$ de l'équation $f(s)=s$. Après calcul on obtient

$$f(s) = \left(1 - \frac{1+\theta}{N} + \frac{1+\theta}{N}s\right)^N.$$

Nous devons donc résoudre l'équation,

$$s = \left(1 - \frac{1+\theta}{N}(1-s)\right)^N.$$

De cette équation, on peut déduire l'inéquation

$$s \leq \exp(-(1+\theta)(1-s)).$$

De cette dernière, on déduit

$$s \leq 1-(1+\theta)(1-s)+(1+\theta)^2(1-s)^2,$$

c'est à dire

$$q \leq 1 - \frac{\theta}{(1+\theta)^2} + o(\theta).$$

Ce qui prouve la première partie du lemme.

Nous allons maintenant étudier la taille moyenne des générations dans le cas où le processus s'éteint. Le théorème A.4 de l'appendice A nous permet de considérer, étant donné son extinction, que le processus évolue comme un processus dont la fonction génératrice est $\hat{f}(s) = q^{-1}f(sq)$. Ce dernier processus s'éteint et son nombre moyen d'enfants est $\hat{f}'(0) = \hat{m} = (1+\theta)q$, qui est bien inférieur à 1. Ainsi, conditionnellement à l'extinction du processus on a,

$$\begin{aligned}\mathbb{E}\,[S|\textit{ extinction }] &= \sum_n \hat{m}^n \\ &= \sum_n ((1+\theta)q)^n \\ &= \frac{1}{1-(1+\theta)q} \\ &\leq 1+\theta.\end{aligned}$$

Ce qui achève la démonstration du lemme 5.2. □

Nous avons maintenant besoin d'évaluer la taille de la frontière d'un arbre de Galton-Watson, conditionnellement à sa survie. Nous considérons donc un arbre aléatoire et la suite de ses frontières $(B_i)_{i\geq 0}$, comme cela a été défini plus haut. A cet effet, nous montrons le lemme suivant,

Lemme 5.3. *Soit une suite* $(|B_k(T)|)_{k\geq 0}$ *de variables aléatoires, définies comme précédemment. Alors pour tout arbre* T*, conditionnellement à sa survie,*

$$|B_k(T)| \geq Bin(kN, (1+\theta)/N) \quad \textit{stochastiquement.} \tag{8}$$

Preuve du lemme 5.3. Afin de montrer ce lemme, nous allons étudier une suite particulière de variables aléatoires et montrer l'inégalité (8) grâce à une méthode de couplage.

Soit la suite de variables aléatoires $(Y_k)_{k\geq 0}$ telle que $Y_k = Y_{k-1} + Z_k - 1$, avec pour tout $k \geq 0$, Z_k est distribué selon la loi $\mathrm{Bin}(N, (1+\theta)/N)$.

On considère maintenant un arbre de Galton-Watson qui ne s'éteint pas. On sait qu'un tel arbre évolue comme un arbre biaisé dans lequel, à chaque génération on forcerait un sommet à avoir au moins un enfant (voir par exemple Lyons et Peres [**29**]). La construction de cet arbre se fait génération par génération et dans une même génération, sommet après sommet. Appelons $(C_k)_{k\geq 0}$ la suite de variables aléatoires représentant la taille des frontières successives de cet arbre biaisé. Les variables C_k sont alors distribuées de la manière suivante,

- Si Z_k est strictement positif, alors $C_k = C_{k-1} + Z_k - 1$;
- sinon, si le sommet considéré est le sommet particulier (celui qui ne s'éteint pas dans la génération considérée), alors $C_k = C_{k-1}$, sinon $C_k = C_{k-1} - 1$.

Alors, pour tout $k \geq 0$ on a $C_k \geq Y_k$ presque sûrement. Mais la loi de distribution de Y_k est $\mathrm{Bin}(kN, (1+\theta)/N) - k$, d'où le lemme 5.3 □

3.2. Analyse de l'algorithme. L'analyse qui a été faite précédemment pour un processus de Galton-Watson va nous servir pour analyser le comportement de l'algorithme. En effet, au temps i, on choisit un sommet v de l'ensemble $B^+_{i-1}(u)$ et on considère l'ensemble noté $O(v)$ de ses enfants. Alors $B^+_i(u) = B^+_{i-1}(u) \backslash \{v\} \cup O(v)$ et comme $|O(v)|$ suit une distribution de loi $\mathrm{Bin}(N, p)$ avec N et p constants, l'algorithme se comporte comme un processus de Galton-Watson. En effet, dans l'algorithme, les sommets sont gérés de telle manière que N et p restent constants. Cette gestion est effectuée grâce à une "migration" de sommets à chaque étape. A cet effet, nous avons créé un ensemble H au début de l'algorithme contenant des sommets qui ne sont pas considérés lors de l'exécution de l'algorithme. A chaque étape, on fait migrer $|O(v)|$ sommets de H vers l'ensemble T de sommets sur lequel on travaille. Ainsi l'ensemble des sommets de travail est de cardinal constant tout au long de l'algorithme et on obtient une loi de reproduction constante $\mathrm{Bin}(N, (1+\theta)/N)$. Le comportement de la procédure est alors similaire au comportement du processus de Galton-Watson étudié dans la section 3.1. Cette similitude permet d'effectuer une analyse de l'algorithme.

Nous allons maintenant montrer que l'algorithme permet, avec grande probabilité, de trouver deux arbres dans $G(F)$ tels que la frontière de chacun d'entre eux est de cardinal au moins $w(n)\sqrt{2n}$. A cette fin, nous allons tout d'abord montrer, grâce au lemme 5.4 qu'il existe presque sûrement un entier n_0 tel que la ${n_0}^{\text{ième}}$ frontière de l'arbre développé atteint la taille $w(n)\sqrt{2n}$ (conditionnellement à la survie de l'arbre). Ensuite, le lemme 5.5 indique que l'algorithme réussit presque sûrement au bout de $1/4\theta^2$ essais, c'est à dire qu'il parvient au bout d'un certain nombre d'essais à développer des arbres suffisamment grands. Enfin, on montrera que l'algorithme n'explore que $1/2\theta^3$ sommets pendant son exécution, avec grande probabilité. Ceci est l'objet du lemme 5.6.

Lemme 5.4. *Etant donné un arbre construit par l'algorithme MAIN, et conditionnellement à sa survie, il existe un entier positif* n_0 *tel que la* ${n_0}^{ième}$ *frontière atteint la taille* $w(n)\sqrt{2n}$ *avec grande probabilité.*

Preuve du lemme 5.4. Soit un arbre construit par l'algorithme MAIN et survivant. Grâce au lemme 5.3 nous pouvons minorer la taille de la $k^{\text{ième}}$ frontière de l'arbre. On en déduit que si on pose $n_0 \geq \lceil 2w(n)n^{3/4} \rceil$, alors la ${n_0}^{\text{ième}}$ frontière est de taille supérieure à $w(n)\sqrt{2n}$ avec grande probabilité. □

Lemme 5.5. *L'algorithme présenté dans la section 2.2 trouve presque sûrement deux sous-ensembles de $G(F)$ de frontière de taille au moins $w(n)\sqrt{2n}$ après au plus $1/4\theta^2$ essais.*

Preuve du lemme 5.5. La probabilité qu'un arbre de Galton-Watson ait une génération de taille supérieure à $w(n)\sqrt{2n}$ est supérieure à sa probabilité de non extinction (en effet, d'après le théorème A.2 un processus qui ne s'éteint pas a des générations dont la taille tend vers l'infini). D'après le lemme 5.2, celle-ci est inférieure à $2\theta + o(\theta)$. L'algorithme présenté dans la section 2.2 développe deux tels arbres de manière indépendante, la probabilité de réussite est donc supérieure à $4\theta^2 + o(\theta^2)$. Si on note $\sharp essais$ le nombre d'exécutions de l'algorithme avant la réussite,

$$\begin{aligned}\mathbb{P}\mathrm{r}\left[\sharp essais \geq \frac{w(n)}{4\theta^2}\right] &\leq \left(1 - 4\theta^2 + o(\theta^2)\right)^{\frac{w(n)}{4\theta^2}} \\ &\leq \exp\left(-3w(n) + o(w(n))\right) \ .\end{aligned}$$

Ce qui achève la preuve. □

Lemme 5.6. *L'algorithme présenté dans la section 2.2 n'explore pas plus de $w(n)/2\theta^3$ sommets avant de trouver deux sous-ensembles de frontière de taille au moins $w(n)\sqrt{2n}$, avec grande probabilité.*

Preuve du lemme 5.6. Après chaque essai, si l'arbre développé n'a pas une frontière suffisamment grande, on ne considère plus les sommets qui viennent d'être parcourus. Ainsi, chaque nouvel essai est effectué sur un graphe qui a perdu environ $2/\theta$ sommets. Afin de déterminer combien de sommets ont été utilisés de cette manière, nous utilisons l'inégalité de Markov. D'après le lemme 5.5, l'algorithme va essayer $w(n)/4\theta^2 = r$ fois, avec grande probabilité, avant de s'arrêter. Durant ces r essais, il va construire $2r$ arbres de tailles respectivement, $T_1, T_1', \ldots, T_r, T_r'$. La somme des tailles de ces arbres est alors la somme de $w(n)/4\theta^2$ variables aléatoires, chacune de moyenne $\mathbb{E}[T] = 2/\theta$. Alors,

$$\mathbb{E}\left[T_1 + T_1' + \ldots + T_r + T_r'\right] = \frac{w(n)}{4\theta^2}\mathbb{E}[T] = \frac{w(n)}{2\theta^3} \ .$$

Et, en utilisant l'inégalité de Markov,

$$\mathbb{P}\mathrm{r}\left[T_1 + T_1' + \ldots + T_m + T_m' \geq t\right] \leq \frac{w(n)}{2\theta^3 t} \ .$$

Si on choisit $t = w(n)/2\theta^3$, alors le lemme 5.6 est prouvé. □

Nous pouvons désormais définir les valeurs des constantes ε et δ. Afin que l'analyse qui vient d'être présentée soit valide, il faut que la moyenne de la loi de distribution $\mathrm{Bin}(N, (1+\theta)/N)$ soit inférieure au nombre moyen de sommets atteignables, c'est à dire,

$$N\frac{1+\theta}{N} \leq \left(\frac{1+\varepsilon}{2n}\right)\left(2n - \frac{w(n)}{2\theta^3}\right),$$

ou encore,

$$4n\theta^4 - 4n\varepsilon\theta^3 + (1+\varepsilon)w(n) \leq 0 \ . \tag{9}$$

La fonction $\theta \mapsto 4n\theta^4 - 4n\varepsilon\theta^3 + (1+\varepsilon)w(n)$ atteint un minimum en $\theta = 3\varepsilon/4$. L'inégalité (9) s'écrit alors,

$$-c\varepsilon^4 n + (1+\varepsilon)w(n) \leq 0,$$

avec c une constante. Cette inégalité est effectivement vérifiée en prenant $\varepsilon \geq w(n)n^{-1/4}$. Alors, sous ces conditions, l'algorithme est en mesure de développer deux arbres de frontière supérieure à $w(n)\sqrt{2n}$. Nous allons maintenant montrer que la présence de tels ensembles permet de conclure au théorème 5.1.

3.3. La probabilité d'un chemin. Nous montrons ici que si deux ensembles $X(u_0)$ et $X(\neg u_0)$, contiennent chacun $w(n)\sqrt{2n}$ sommets et sont indépendants, alors dans un graphe tel que $G(F)$, il existe presque sûrement un chemin entre eux. La condition d'indépendance est garantie du fait que nous considérons les frontières des arbres dont on n'a pas eploré le voisinage.

Soit u_0 un sommet de $G(F)$. On suppose que $|X(u_0)| \geq w(n)\sqrt{2n}$ et $|X(\neg u_0)| \geq w(n)\sqrt{2n}$. Soit u un sommet de $X(u_0)$, alors, par construction, $\neg u$ appartient à l'ensemble $Y(\neg u_0)$. Donc, s'il existe une arête entre $u \in X(u_0)$ et $v \in Y(\neg u_0)$, il en existe aussi une entre $\neg u$ et $\neg v$. Chaque paire d'arête existe avec probabilité p, donc la probabilité qu'il n'existe pas de telle paire (u, v) est,

$$(1-p)^{2nw^2(n)} = \exp(-w^2(n) + O(1/n)).$$

On en déduit alors qu'il existe presque sûrement une arête entre $X(u_0)$ et $Y(\neg u_0)$. La même méthode permet de conclure à l'existence d'une arête entre $X(\neg u_0)$ et $Y(u_0)$. Cela conclut la démonstration du théorème 5.1.

CHAPITRE 6

Sur les fluctuations de la composante géante

1. Introduction

Nous nous intéressons maintenant aux composantes fortement connexes dans un graphe aléatoire orienté. On considère le graphe $D_{n,p}$ contenant n sommets $x_1, \ldots, x_n$ et où chaque arc est présent avec probabilité p. Il a été montré par Karp [**26**] que si n est grand et np est égal à une constante c supérieure à 1, alors avec grande probabilité toutes les composantes connexes fortes contiennent peu de sommets, sauf une (que l'on appellera *composante connexe géante*) qui contient environ $\Theta^2 n$ sommets, où Θ est la racine de l'équation $1 - x - e^{-cx} = 0$ dans $]0, 1[$. De plus Karp a montré que la taille de cette composante connexe géante fluctue d'au plus $w(n)\sqrt{n \log n}$ sommets, où $w(n)$ est une fonction croissante non bornée. Karp a ainsi montré le théorème suivant,

Théorème 6.1 (Karp)**.** *Soit $w(n)$ une fonction croissante non bornée et c une constante strictement supérieure à 1. Il existe une constante M telle que, avec grande probabilité, un graphe aléatoire orienté $D_{n,c/n}$ a exactement une composante fortement connexe contenant plus de $M \log n$ sommets, et la taille de cette composante diffère de $\Theta^2 n$ sommets par au plus $w(n)\sqrt{n \log n}$ sommets.*

Dans ce chapitre nous allons exposer la preuve de Karp [**26**] de ce résultat qui utilise des résultats concernant la taille de processus de branchement poissonniens. Ensuite nous montrerons comment on peut améliorer ce résultat en décomposant des arbres de Galton-Watson, pour obtenir le théorème suivant,

Théorème 6.2 (Verhoeven)**.** *Soit $w(n)$ une fonction croissante non bornée. Soit c une constante supérieure à 1. Il existe une constante K telle que, presque sûrement, un graphe aléatoire orienté $D_{n,c/n}$ a exactement une composante fortement connexe contenant plus de $K \log n$ sommets et le nombre de sommets dans cette composante appartient à $[\Theta^2 n - w(n)\sqrt{n}, \Theta^2 n + w(n)\sqrt{n}]$.*

Dans la suite, nous supposons $c > 1$ et donc, comme nous allons le voir, le graphe contient presque sûrement une composante fortement connexe géante.

2. La structure de la clôture transitive

Nous allons montrer le théorème 6.1 en le décomposant. Des détails concernant cette démonstration peuvent être trouvés dans [**26**]. Tout d'abord, nous montrons un résultat sur la taille de la composante géante fortement connexe. Certaines démonstrations présentées dans les chapitres 4, 5 et l'appendice C visant à déterminer des ensembles atteignables à partir d'un sommet sont similaires à la démonstration présentée ici. Comme dans ces chapitres, pour un sommet u, l'ensemble $X(u)$ représente l'ensemble des sommets atteignables à partir de u. Soit u un sommet quelconque. On a le théorème suivant

Théorème 6.3 (Karp)**.** *Soit a une constante positive et soit B une constante plus grande que $(a+1)c(c-1)^{-2}$. Soit $w(n)$ une fonction croissante non bornée. Alors, pour n suffisamment grand, $\Pr[|X(u)| \notin [0, B \ln n] \cup [\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]] < n^{-a}$.*

Preuve du théorème 6.3. Nous allons considérer le processus suivant pour construire l'ensemble $X(u)$. Le processus va générer une suite $((A_i, B_i))_{i\geq 0}$ avec $A_i \subseteq \{1,\ldots,n\}$ et $|A_i| = i$. L'ensemble B_i est constitué des sommets qui ont été atteints pendant les i premières itérations du processus et l'ensemble A_i est constitué des sommets qui ont été explorés pendant les i premières itérations du processus. On a $A_0 = \emptyset$ et $B_0 = \{u\}$ et le couple (A_{i+1}, B_{i+1}) est obtenu à partir de (A_i, B_i) en effectuant : $A_{i+1} = A_i \cup \{v\}$ et $B_{i+1} = B_i \cup succ(v)$, où v est un sommet choisie aléatoirement parmi les éléments de $B_i \backslash A_i$ et $succ(v)$ est l'ensemble des sommets w tels qu'il existe une arête de v vers w. Le processus s'arrête lorsque $A_t = B_t$, c'est à dire quand tous les sommets atteignables à partir de u ont été explorés. Alors $A_t = B_t = X(u)$.

Analysons maintenant ce processus. A chaque itération, chaque sommet non encore atteint a la probabilité p d'être atteint. On peut donc décrire le comportement du processus en termes de probabilités. Soit $\mathrm{Bin}(n,p)$ la distribution binomiale de paramètres n et p. La distribution conditionnelle de $|B_{i+1} \backslash B_i|$ sachant $B_0, \ldots, B_i$ est $\mathrm{Bin}(n - |B_i|, p)$ (ici B_{i+1} est défini seulement si B_i est défini et $|B_i| > i$). Il s'ensuit que l'on peut coupler la suite B_i avec une suite Z_i telle que $Z_0 = 1$ et $Z_{i+1} = Z_i + \mathrm{Bin}(n - Z_i, p)$. Les deux suite $(B_i)_{i\geq 0}$ et $(Z_i)_{i\geq 0}$ ont la même distribution, mais la suite $(B_i)_{i\geq 0}$ s'arrête dès qu'un t est tel que $|B_t| = t$ tandis que $(Z_i)_{i\geq 0}$ continue. Il 'ensuit que $|X(u)|$ a la même distribution que $\min\{t : Z_t = t\}$ et $\mathbb{P}\mathrm{r}[|X(1)| = t] \leq \mathbb{P}\mathrm{r}[Z_t = t]$.

Il est facile de montrer par récurrence que $Z_t - 1$ est distribuée comme $\mathrm{Bin}(n - 1, 1 - (1-p)^t)$. On va utiliser les deux bornes suivantes concernant la queue de la distribution binomiale [**34**]. Soit X une variable aléatoire dont la distribution est $\mathrm{Bin}(n,p)$.

$$\text{Pour tout } \beta \in \mathbb{R}_+^*,\ \mathbb{P}\mathrm{r}[X > \beta np] \leq \left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{np}, \tag{10}$$

$$\text{pour tout } \gamma \in [0,1],\ \mathbb{P}\mathrm{r}[X < (1-\gamma)np] < e^{-\gamma^2 np/2}. \tag{11}$$

On définit maintenant la suite de variables aléatoires $(W_i)_{i\geq 0}$ telle que $W_0 = 1$ et $W_{i+1} = W_i + \mathrm{Bin}(n-t, p)$. On montre par récurrence que, pour tout $j \leq t$, et tout i, $\mathbb{P}\mathrm{r}[Z_j \leq i] \leq \mathbb{P}\mathrm{r}[W_j \leq i]$. Donc, $\mathbb{P}\mathrm{r}[|X(1)| = t] \leq \mathbb{P}\mathrm{r}[W_t = t]$. Or $W_t - 1$ est distribué comme $\mathrm{Bin}(t(n-t), p)$ et d'après (10), $\mathbb{P}\mathrm{r}[W_t \leq t] \leq \exp(-((c-1)t - ct^2/n)^2/ct)$. Après calcul, on voit alors que pour $B > (a+1)c/(c-1)^2$, et pour n suffisamment grand, $\mathbb{P}\mathrm{r}[W_t \leq t] \leq n^{-a-1}$ pour tout $t \in [B \ln n, (c-1)n/3]$.

Considérons maintenant le cas $t \geq (c-1)n/3$. On se souvient que $Z_t - 1$ a comme loi de distribution $\mathrm{Bin}(n-1, 1-(1-c/n)^{xn})$, avec $x = t/n$. En utilisant le fait que

i. pour tout $y \in \mathbb{R}$, $\ln(1-y) < -y$ et
ii. pour $0 < y < .69$, $\ln(1-y) > -y - y^2$.

On obtient alors

$$\mathbb{P}\mathrm{r}[Z_t \leq t] \leq \exp\left(-n\frac{(1 - e^{-(cx + c^2x/n) - x})^2}{2(1 - e^{-cx})}\right)$$

et un bref calcul donne alors, avec $t = xn$, $x \leq \Theta - w(n)n^{-1/2}$ et n suffisamment grand, $\mathbb{P}\mathrm{r}[Z_t \leq t] < n^{-a-1}$, où $w(n)$ est une fonction croissante non bornée.

Un calcul similaire mais utilisant l'inégalité (11) et avec $t = xn$, $x > \Theta + w(n)n^{-1/2}$ et n suffisamment grand, $\mathbb{P}\mathrm{r}[Z_t \geq t] < n^{-a-1}$.

On a donc montré que pour n suffisamment et pour tout t en dehors des intervalles $[0, B \ln n]$ et $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]$, $\mathbb{P}\mathrm{r}[Z_t = t] < n^{-a-1}$. Il s'ensuit que quand n est suffisamment grand, la probabilité que $|X(1)|$ soit en dehors l'union de ces intervalles est inférieure à n^{-a-1}. □

Dans la suite nous dirons qu'un ensemble est "petit" lorsque sa taille est dans l'intervalle $[0, B\ln n]$ et "grand" quand la taille de l'ensemble appartient à l'intervalle $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]$. Pour deux sommets u et v, on notera uRv quand le sommet v est atteignable à partir du sommet u.

Une démonstration analogue à celle du lemme 5.2 du chapitre 5 permet de montrer que la probabilité de $X(u)$ soit petit est $1-\Theta$ et dans ce cas l'espérance de la taille de $X(u)$ est $1/(1-c(1-\Theta))$.

Théorème 6.4. *Soit un graphe dirigé aléatoire $D_{n,c/n}$ avec $c > 1$. Alors, avec probabilité qui tend vers 1, pour tout u et v, si $X(u)$ est grand et $Y(v)$ est grand, alors uRv.*

Preuve du théorème 6.4. Il n'y a que n^2 choix possibles pour le couple (u, v), il suffit donc de montrer que pour tout couple (u, v) on a $\mathbb{P}\mathrm{r}[X(u)$ est grand $\cap$ $Y(v)$ est grand $\cap u\neg Rv] = o(n^{-2})$. Cette probabilité est majorée par la probabilité $\mathbb{P}\mathrm{r}[X(u)$ est grand $\cap\, Y(v)$ est grand $|u\neg Rv]$, nous allons donc montrer que cette probabilité conditionnelle est $o(n^{-2})$. On considère le processus permettant de construire l'ensemble $X(u)$. Sachant que $u\neg Rv$, il n'existe pas d'arc d'un sommet de $X(u)$ vers un sommet de $Y(v)$. Ainsi, lors du processus, le nombre d'arc sortant d'un sommet est stochastiquement plus petit que $\mathrm{Bin}(n - |Y(v)|, c/n)$. On rappelle qu'une variable aléatoire X est stochastiquement plus petite qu'une variable aléatoire Y $(X \prec Y)$ si et seulement si, pour tout $t \in \mathbb{R}$, on a $\mathbb{P}\mathrm{r}[X < t] \leq \mathbb{P}\mathrm{r}[Y < t]$. Pour tout N, la probabilité que le processus atteigne $N+1$ sommets est majorée par la probabilité qu'au moins N sommets soient touchés durant les N premières itérations du processus. Mais, la distribution de probabilité du nombre de sommets touchés durant les N premières itérations est stochastiquement plus petite que $\mathrm{Bin}(N(n - |Y(v)|), c/n)$. Si on choisit $N = \Theta n - w(n)\sqrt{n}$, et en remarquant que $Y(v) \geq \Theta n - w(n)\sqrt{n}$ et $c(1-\Theta) < 1$ et en appliquant la borne 10, on trouve que $\mathbb{P}\mathrm{r}[X(u)$ est grand$\cap Y(v)$ est grand$|u\neg Rv]$ est borné par une fonction qui tend de façon exponentielle vers 0. □

On notera $A(u)$ l'événement que $X(u)$ soit "grand" et $B(v)$ l'événement que $Y(v)$ soit grand. On montrera que ces $2n$ évènements sont positivement corrélés deux à deux. On déduit d'un résultat général donné par Harris [**24**] concernant la corrélation positive des propriétés monotones que deux quelconques de ces évènements soient positivement corrélés. Ainsi, par exemple, quels que soient les sommets u et v,

$$\mathbb{P}\mathrm{r}[A(u)|\neg A(v)] \leq \mathbb{P}\mathrm{r}[A(u)] \leq \mathbb{P}\mathrm{r}[A(u)|A(v)].$$

Le lemme suivant montre que la corrélation positive n'est pas très forte.

Lemme 6.5. *Soient E et F deux évènements distincts appartenant à l'ensemble $\{A(u), u = 1, 2, \ldots, n\} \cup \{B(v), v = 1, 2, \ldots, n\}$. Alors*

$$\mathbb{P}\mathrm{r}[E|\neg F] \geq \mathbb{P}\mathrm{r}[E] - O\left(\frac{\log n}{n}\right).$$

Preuve du lemme 6.5. Considérons par exemple le cas où $E = A(u)$ et $F = A(v)$, $u \neq v$. Nous allons montrer que, pour chaque petit ensemble X ne contenant pas u, $\mathbb{P}\mathrm{r}[A(u)|X(v) = X] \geq \mathbb{P}\mathrm{r}[A(u)] - O(\log n/n)|X|$. Considérons une expérience pour déterminer, à l'aide d'un processus de *fanning-out*, si $X(u)$ est grand. Cette expérience s'interrompt dès que $M\ln n$ sommets sont atteints et l'expérience est indépendante du fait que $X(v) = X$, à moins qu'un sommet de X soit touché pendant l'exploration. Mais la probabilité qu'un ensemble aléatoire $X(u)/u$ de taille au plus $M \ln n$ intersecte X est au plus $M \ln n|X|/n$. Sachant que $X(v)$ est petit, l'espérance de la taille de $X(v)$ est majorée par une constante, et la probabilité que u soit dans $X(v)$ est $O(1/n)$. Les autres cas se traitent de façon similaire. Le lemme est alors montré. $\square$

Corollaire 6.6. *Soient E et F deux évènements distincts appartenant à l'ensemble $\{A(u), u = 1, 2, \ldots, n\} \cup \{B(v), v = 1, 2, \ldots, n\}$. Alors $\mathbb{P}\mathrm{r}[E|F] \leq \mathbb{P}\mathrm{r}[E] + O(\log n/n)$.*

Preuve du corollaire 6.6. Le résultat suit directement du lemme 6.5 et en utilisant l'inégalité $\mathbb{P}\mathrm{r}[E] = \mathbb{P}\mathrm{r}[E|F]\mathbb{P}\mathrm{r}[F] + \mathbb{P}\mathrm{r}[E|\neg F]\mathbb{P}\mathrm{r}[\neg F]$ et le fait que $\mathbb{P}\mathrm{r}[E]$, $\mathbb{P}\mathrm{r}[F]$ et $\mathbb{P}\mathrm{r}[\neg F]$ tendent tous vers des constantes positives quand n tend vers l'infini. $\square$

Corollaire 6.7. *Soient $LARGEOUT = |\{u : X(u) \text{ est grand}\}|$, $LARGEIN = |\{v : Y(v) \text{ est grand}\}|$ et $LARGE = |\{u : X(u) \text{ est grand et } Y(u) \text{ est grand}\}|$. Soit $w(n)$ une fonction croissante non bornée. Alors, avec probabilité tendant vers 1, $|LARGEOUT - \Theta n| < w(n)\sqrt{n \log n}$, $|LARGEIN - \Theta n| < w(n)\sqrt{n \log n}$ et $|LARGE - \Theta^2 n| < w(n)\sqrt{n \log n}$.*

Preuve du corollaire 6.7. Les preuves des trois résultats sont similaires. Montrons le résultat concernant la variable aléatoire $LARGEOUT$ qui est égale à $\sum_{u=1}^{n} A(u)$. D'après le lemme 5.2, $\mathbb{E}[A(1)]$ tend vers Θ et il est possible de montrer que $\mathbb{E}[A(1)] = \Theta + O(n^{-1})$. Alors, $\mathbb{E}[LARGEOUT] = \Theta n + O(1)$ et $\mathbb{E}[(LARGEOUT)^2] = \mathbb{E}[LARGEOUT] + n(n-1)\mathbb{E}[A(u)A(v)]$, où u et v sont deux sommets distincts quelconques. En appliquant le corollaire 6.6, on a

$$\mathbb{E}[A(u)A(v)] = \mathbb{E}[A(u)]\mathbb{E}[A(v)|A(u)] = \mathbb{E}[A(u)](\mathbb{E}[A(v)] + O(\log n/n)).$$

Il s'ensuit que $LARGEOUT$ appartient à $O(n \log n)$ et la borne attendue découle de l'inégalité de Chebyshev. $\square$

Le théorème 6.3 découle alors du théorème 6.4 et du corollaire 6.7.

Preuve du théorème 6.3. Toute composante connexe forte contenant un sommet qui n'est pas dans $LARGE$ est de taille inférieure à $M \log n$. D'après le théorème 6.4, avec probabilité tendant vers 1, pour toute paire de sommets $\{u, v\}$ appartenant à $LARGE$, uRv. Dans ce cas l'ensemble $LARGE$ est une composante connexe forte. Le résultat découle alors du corollaire 6.7. $\square$

3. Une amélioration du théorème de Karp

Nous allons maintenant montrer comment transformer la preuve du lemme 6.5 de façon à obtenir une borne permettant d'en déduire le théorème 6.2. Pour cela nous montrons le lemme suivant qui est plus précis que le lemme 6.5.

Lemme 6.8. *Soient E et F deux évènements distincts de l'ensemble $\{A(u), u = 1, 2, \ldots, n\} \cup \{B(v), v = 1, 2, \ldots, n\}$ définis ci-dessus. Alors*

$$\mathbb{P}\mathrm{r}[E|\neg F] \geq \mathbb{P}\mathrm{r}[E] - O\left(\frac{1}{n}\right).$$

Pour montrer ce lemme, nous allons avoir besoin du lemme outil 6.9,

Lemme 6.9. *Soit T^* le sous arbre infini d'un arbre de Galton-Watson dont la loi de reproduction est une loi de Poisson de paramètre λ ($\lambda \geq 1$) et dont la probabilité d'extinction est q. On notera $(Z_n^*)_{n \geq 0}$ la suite des tailles des générations de T^*. Si on pose $\overline{q} = 1 - q$ et $\tau = \max\{k \in \mathbb{N} \ : \ Z_k^* = 1\}$, alors,*

$$\mathbb{P}\mathrm{r}[\tau = k] = \left(\frac{1}{\overline{q}} - \lambda e^{-\lambda \overline{q}} - \frac{e^{-\lambda \overline{q}}}{\overline{q}}\right)\left(\lambda e^{-\lambda \overline{q}}\right)^k.$$

Nous commençons donc par montrer ce lemme.

Preuve du lemme 6.9. τ est le dernier indice tel que la taille de la génération Z^*_τ soit égale à 1. On écrit donc, on remarquant qu'un arbre de Galton-Watson est un processus Markovien,

$$\begin{aligned}
\mathbb{P}\mathrm{r}[\tau = k] &= \mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2, Z^*_k = 1, \ldots, Z^*_0 = 1] \\
&= \mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2 | Z^*_k = 1, \ldots, Z^*_0 = 1]\mathbb{P}\mathrm{r}[Z^*_k = 1, \ldots, Z^*_0 = 1] \\
&= \mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2 | Z^*_k = 1]\mathbb{P}\mathrm{r}[Z^*_k = 1, \ldots, Z^*_0 = 1] \\
&= \mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2 | Z^*_k = 1]\mathbb{P}\mathrm{r}[Z^*_k = 1 | Z^*_{k-1} = 1] \ldots \mathbb{P}\mathrm{r}[Z^*_1 = 1 | Z^*_0 = 1] \\
&= \mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2 | Z^*_k = 1]\mathbb{P}\mathrm{r}[Z^*_1 = 1 | Z^*_0 = 1]^k \qquad (12)
\end{aligned}$$

Or la fonction génératrice de la loi de Poisson de paramètre λ est

$$f(s) = e^{-\lambda + \lambda s},$$

donc avec ce qui a été dit précédemment, la fonction génératrice du sous arbre infini est

$$\begin{aligned}
f^*(s) &= \frac{e^{-\lambda\overline{q} + \lambda\overline{q}s} - q}{\overline{q}} \\
&= \frac{e^{-\lambda\overline{q}}}{\overline{q}} \sum_{k\geq 0} \frac{(\lambda\overline{q})^k}{k!} s^k - \frac{q}{\overline{q}}.
\end{aligned}$$

On en déduit donc

$$\mathbb{P}\mathrm{r}[Z^*_1 = 1 | Z^*_0 = 1] = \lambda e^{-\lambda\overline{q}}$$

et

$$\mathbb{P}\mathrm{r}[Z^*_{k+1} \geq 2 | Z^*_k = 1] = \frac{e^{-\lambda\overline{q}}}{\overline{q}} \sum_{k\geq 2} \frac{(\lambda\overline{q})^k}{k!} = \frac{1}{\overline{q}} - \lambda e^{-\lambda\overline{q}} - \frac{e^{-\lambda\overline{q}}}{\overline{q}}.$$

En remplaçant dans (12), on obtient le résultat désiré. □

Nous pouvons maintenant donner une démonstration simple du lemme 6.8 qui nous permettra d'obtenir l'intervalle des fluctuations de la taille de la composante fortement connexe géante. Pour cela on va tenter de déterminer dans quelle mesure la connaissance de $\neg F$ influe sur l'événement E. À cette fin, nous supposons que $E = A(u)$ et $F = A(v)$ et on va développer l'ensemble des sommets atteignables à partir de u. Cet ensemble se comporte comme un arbre de Galton-Watson surcritique pendant un certain nombre de générations (pour une justification, on peut consulter le chapitre 5). Un tel arbre comporte un sous-arbre dont les branches sont "infinies" et dont l'existence garantie l'obtention d'un "grand" ensemble. Aussi on va estimer la probabilité qu'une de ces branches rencontre l'ensemble des sommets atteignables à partir de v car dans ce cas, la taille de la branche est alors limitée. Par ailleurs on montre qu'il est peu probable que les deux ensembles s'intersectent en deux points, aussi, si le sous-arbre contient deux branches la probabilité que celui-ci s'éteigne est $o(1/n)$.

Preuve du lemme 6.8. On considère le cas où $E = A(u)$ et $F = A(v)$, $u \neq v$, les autres cas sont similaires. Pour cela on va étudier un procédé construisant l'arbre $X(u)$ issu de u génération après génération. On notera $X(v)$ l'arbre issu de v. Nous avons vu qu'un tel arbre est soit de taille inférieure à $M \log n$, M constante (et $\mathbb{E}[X(v)] = C$, C constante), soit de taille linéaire, en négligeant une probabilité $o(1/n)$. Il nous suffit donc de construire l'arbre issu de u sans dépasser la taille $M \log n$.

Nous allons utiliser un couplage entre un processus de construction d'arbre dans le graphe (différend de celui de Karp) avec un processus de Galton-Watson

T_u défini par une distribution de reproduction distribuée suivant une loi de Poisson de paramètre c (le processus est donc sur-critique). Le processus $(Y_k(u))_{k\geq 0}$ du graphe est défini de la façon suivante : soit u un sommet fixé. Alors, pour chaque $k \geq 0$, $Y_k(u)$ est le nombre de sommets à distance k de u. La représentation de ce processus par un processus de Galton-Watson n'est pas rigoureusement exacte, mais l'approximation nous semble suffisante ici du fait que nous ne considérons que $O(\log n)$ sommets.

On cherche à étudier l'arbre issu de u conditionnellement à ce que les sommets d'un certain ensemble V soient inactifs. Cela revient à étudier un arbre de Galton-Watson non conditionné dans lequel on enlève $|V|$ sommets, chacun avec la probabilité $|V|/n$.

On notera f la fonction génératrice du processus de Galton-Watson et q sa probabilité d'extinction. D'après Athreya et Ney [**5**, pp. 47-53] il existe un sous-arbre T_u^* dont les sommets sont ceux de T_u qui ont une descendance infinie (la démonstration de ce résultat est présentée dans l'appendice A). On peut associer à cet arbre un autre processus de Galton-Watson dont la probabilité d'extinction est évidemment nulle et la fonction génératrice est $f^*(s) = (f((1-q)s+q) - q)/(1-q)$. On notera Z_k (respectivement Z_k^*) la taille de la $k^{\text{ième}}$ génération de T_u (respectivement T_u^*). On peut remarquer que la suite Z_k^* est croissante. Enfin, on notera $T_{u,k}^*$ le sous arbre fini de T_u^* contenant les k premières générations de T_u^*.

Soit τ défini de la façon suivante :

$$\tau = \max\{k \in \mathbb{N} \ : \ Z_k^* = 1\}.$$

Si pour $k \leq \tau$, $T_{u,k}^* \cap X(v) \neq \emptyset$, alors, on sait que l'arbre issu de u est "petit". Par ailleurs, on sait que si toute une génération de T_u^* appartient à $X(v)$, alors l'arbre issu de u est "petit". Ainsi, on a

$$\mathbb{P}\mathrm{r}[E|\neg F] \geq \mathbb{P}\mathrm{r}[E] - \mathbb{P}\mathrm{r}[|X(v) \cap T_{u,\tau}^*| \geq 1] - \mathbb{P}\mathrm{r}[|X(v) \cap X(u)| \geq 2].$$

Sachant que $X(u)$ est au plus de cardinal $M \log n$, on voit donc que $|X(v) \cap X(u)|$ est stochastiquement dominé par la loi binomiale $\mathrm{Bin}(M \log n, M \log n/n)$ que l'on peut approximer à l'aide de la loi de Poisson de paramètre $M^2 \log^2 n/n$. Donc on a

$$\mathbb{P}\mathrm{r}[|X(v) \cap X(u)| \geq 2] = O\left(\frac{\log^4 n}{n^2}\right).$$

Évaluons maintenant la probabilité $\mathbb{P}\mathrm{r}[|X(v) \cap T_{u,\tau}^*| \geq 1]$. Grâce au lemme 6.9, on peut écrire

$$\begin{aligned}
\mathbb{P}\mathrm{r}[|X(v) \cap T_{u,\tau}^*| \geq 1] &= \sum_{k=0}^{\infty} \mathbb{P}\mathrm{r}[\tau = k] Ck\frac{1}{n} \\
&= \left(\frac{1}{\overline{q}} - \lambda e^{-\lambda\overline{q}} - \frac{e^{-\lambda\overline{q}}}{\overline{q}}\right) C\frac{1}{n} \sum_{k=0}^{\infty} k\left(\lambda e^{-\lambda\overline{q}}\right)^k \\
&= \left(\frac{1}{\overline{q}} - \lambda e^{-\lambda\overline{q}} - \frac{e^{-\lambda\overline{q}}}{\overline{q}}\right) C\frac{1}{n} \left(\frac{1}{\ln(\lambda e^{-\lambda\overline{q}})}(1-\lambda e^{-\lambda\overline{q}})^{-2} - (1-\lambda e^{-\lambda\overline{q}})^{-1}\right) \\
&= O\left(\frac{1}{n}\right)
\end{aligned}$$

car λ, q et $\overline{q}$ sont des constantes.

On en conclut alors

$$\mathbb{P}\mathrm{r}[E|\neg F] \geq \mathbb{P}\mathrm{r}[E] - O(n^{-1}).$$

Les autres cas se traitent de manière similaire. $\square$

Le théorème 6.2 se déduit du lemme 6.8. La démonstration du théorème 6.2 à partir du lemme 6.8 est tout à fait analogue à la démonstration de Karp du théorème 6.1 à partir du lemme 6.5.

Conclusion

On a vu dans cette thèse comment le problème du seuil de 2-SAT a pu être résolu grâce à l'étude du graphe orienté associé à une formule. Toutefois de nombreux autres problèmes se posent concernant 2-SAT. Ainsi on peut étudier le nombre d'assignement d'une formule sous-critique. Sharell [**35**] a pu montrer un résultat de concentration, mais la valeur autour de laquelle cette concentration a lieu est toujours inconnue. D'autre part, Boufkhad et Dubois [**13**] ont montré dans un cadre plus général que pour tout $r \geq 2$, une formule r-SAT est soit satisfaisable avec un nombre de modèles exponentiel dans le nombre de variables, soit insatisfaisable. Ce problème est d'une grande importance en physique statistique òu il a été étudié par Monasson et Zecchina [**31, 32**] et où le logarithme du nombre de solutions d'une formule 2-SAT est l'entropie minimale du système physique correspondant. D'autre part, il semblerait qu'il n'existe aucune étude concernant le nombre de solutions d'une formule sur-critique conditionnée à être satisfaisable. A-t'on la encore un résultat de concentration ?

Par ailleurs, on a pu se rendre compte de l'utilité des processus de branchements de type Galton-Watson dans l'étude de diverses propriétés des composantes fortement connexes dans différents graphes aléatoires orientés. On a vu que grâce à des résultats relativement simples concernant ces processus on obtient des résultats concernant la taille et les fluctuations de la composante fortement connexe géante en particuliers. Pourrait-on utiliser de tels processus pour retrouver les résultats connus sur la composante connexe géante dans le graphe aléatoire non orienté ? Par exemple Barraez, Boucheron et Fernandez de la Vega [**6**] ont montré que les déviations de la taille de cette composante converge en distribution vers une loi Normale en étudiant un processus tel que celui présenté dans les chapitres 5 et 6. A-t'on un résultat analogue pour la composante géante des graphes orientés ?

Bibliographie

[1] D. Achlioptas. Setting two variables at a time yields a new lower bound for random 3-sat. In *32nd Annual ACM Symposium on Theory of Computing*, pages 28–37, New-York, 2000. ACM.

[2] M. Ajtai, J. Komlòs, and E. Szemerédi. The longest path in a random graph. *Combinatorica*, 1(1):1–12, 1981.

[3] N. Alon and J. Spencer. *The Probabilistic Method.* John Wiley & Sons, 1992.

[4] B. Aspvall, M. F. Plass, and R. E. Tarjan. A linear-time algorithm for testing the truth of certain quantified boolean formulas. *Information Processing Letters*, 8:121–123, 1979.

[5] K. B. Athreya and P. E. Ney. *Branching Processes.* Springer Verlag, New York, 1972.

[6] D. Barraez, S. Boucheron, and W. Fernandez de la Vega. On the fluctuations of the giant component. *Combinatorics Probability and Computing*, 9:287–304, July 2000.

[7] B. Bollobás, C. Borgs, J. Chayes, J. Han Kim, and D. B. Wilson. The scaling window of the 2-sat transition. *Rand. Struct. Alg.*, 2001.

[8] M. Chao and J. Franco. Probabilistic analysis of two heuristics for the 3-satisfiability problem. *SIAM J. Comput.*, 15(4):1106–1118, 1986.

[9] M. Chao and J. Franco. Probabilistic analysis of a generalization of the unit-clause literal selection heuristic for the k-satisfiability problem. *Inform. Sci.*, 51(3):289–314, 1990.

[10] V. Chvátal and B. Reed. Micks gets some (the odds are on his side). In *Proceedings 33rd FOCS*, pages 620–627, 1992.

[11] W. Fernandez de la Vega. Random 2-sat: Results and problems. to appear in Theorical Computer Science.

[12] W. Fernandez de la Vega. On random 2-sat (revised version). manuscript, 1992.

[13] O. Dubois and Y. Boufkhad. A general upper bound for the satisfiability threshold of random r-sat formulae. *J. Algorithms*, 24(2):395–420, 1997.

[14] O. Dubois, Y. Boufkhad, and J. Mandler. Typical random 3-sat formulae and the satisfiability threshold. In *11th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 126–127, New-York, 2000. ACM.

[15] P. Erdös. Graph theory and probability. *Canad. J. Math.*, 11:34–38, 1959.

[16] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17–61, 1960.

[17] P. Erdös and A. Rényi. On the evolution of random graphs. *Bull. Inst. Int. Statis. Tokyo*, 38:343–347, 1961.

[18] E. Friedgut. Sharp thresholds of graph properties, and the k-sat problem. *J. Amer. Math. Soc.*, 12:1017–1054, 1999.

[19] A. Frieze and S. Suen. Analysis of two simple heuristics on a random instance of k-sat. *Journal of Algorithmics*, 20:312–355, 1996.

[20] A. Goerdt. A threshold for unsatisfiability. In J.M. Havel and V. Koubek, editors, *Proceedings MFCS*, pages 264–274. LNCS 692, 1992.

[21] A. Goerdt. A remark on random 2-SAT. *Discrete Applied Mathematics*, 96/97:107–110, 1999.

[22] M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors. *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springr-Verlag, 1998.

[23] T. E. Harris. Branching processes. *AMS*, 19:474–494, 1948.

[24] T. E. Harris. *The Theory of Branching Processes*. Springer Verlag, New York, 1963.

[25] B. Hayes. Can't get no satisfaction. *American Scientist*, 85:108–112, March-April 1997.

[26] R. M. Karp. The transitive closure of a random digraph. *Random Structures and Algorithms*, pages 73–93, 1990.

[27] D. G. Kendall. Branching processes since 1873. *Journal of the London Mathematical Society*, 41:385–406, 1966.

[28] L. M. Kirousis, E. Kranakis, D. Krizanc, and Y. C. Stamatiou. Approximating the unsatisfiability threshold of random formulas. *Random Structures and Algorithms*, pages 253–269, 1998.

[29] R. Lyons and Y. Peres. Probability on trees and networks. Site internet : http://php.indiana.edu/˜rdlyons/prbtree/prbtree.html, 1999.

[30] A. El Maftouhi and W. Fernandez de la Vega. On random 3-sat. *Combinatoric, probability and computing*, 4:189–195, 1995.

[31] R. Monasson and R. Zecchina. The entropy of the k-satisfiability problem. *Phys. Rev. Lett.*, 76(3881), 1996.

[32] R. Monasson and R. Zecchina. Statistical mechanics of the random k-sat model. *Phys. Rev. E*, 56(1357), 1997.

[33] C. H. Papadimitriou. *Computational Complexity*. Addison-Wesley, 1994.

[34] P. Raghavan. Probabilistic construction of deterministic algorithms: approximating packing integer programs. In *Proceedings of the 27th Annual IEEE Symp. on Foundations of Computer Science*, pages 10–18, 1986.

[35] A. Sharell. Concentration to the number of solutions to a rnadom 2-sat formula. manuscript, 2000.

[36] R. Tarjan. Depth-rst search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[37] Y. Verhoeven. Random 2sat and unsatisfiability. *Information Processing Letters*, 72(3-4):119–124, 1999.

Index

APPENDICE A

Le processus de Galton-Watson

Nous présentons ici une introduction aux *processus de branchement* particuliers que sont les processus de Galton-Watson. Cette introduction permet de montrer une propriété particulière de ces processus, pour de plus amples informations, le lecteur pourra consulter les ouvrages de Harris [**24**], Athreya et Ney [**5**], Lyons et Peres [**29**] ou encore Devroye [**22**].

1. Le modèle de Galton-Watson

Vers 1873, Galton et Watson introduire un modèle afin d'étudier la pérennité des noms des nobles anglais (on pourra à ce propos consulter l'étude historique de Kendall [**27**]). Leur modèle qui est aujourd'hui connu sous le nom de *processus de Galton-Watson* où encore *arbre de Galton-Watson* est très simple : une population commence avec un parent, la racine. Cette racine a Z_1 enfants, où Z_1 a une distribution fixée (la *distribution de reproduction*). Nous noterons Z une variable aléatoire distribuée comme Z_1. On pose

$$p_i = \mathbb{P}\mathrm{r}[Z = i],\ i \geq 0.$$

Chaque enfant se reproduit alors suivant la même distribution, et ainsi de suite. On construit alors un arbre aléatoire, l'arbre de Galton-Watson et un processus aléatoire, le processus de Galton-Watson. On notera Z_i le nombre de descendants à la $i^{\text{ème}}$ génération, et $Z_0 = 1$. Le processus peut alors évoluer de deux façons, soit la population survit à tout jamais ($Z_i > 0$ pour tout $i \in \mathbb{N}$), ou alors le processus s'éteint au bout d'un temps fini. Nous allons maintenant étudier l'évolution du processus.

2. L'évolution du processus de Galton-Watson

Afin d'analyser le processus décrit précédemment, il est pratique d'utiliser la fonction génératrice

$$f(s) = \sum_{k=0}^{\infty} p_k s^k = \mathbb{E}[s^{Z_1}],\ s \in [0,1].$$

Cette fonction est strictement convexe (si $p_1 \neq 1$) et croit de p_0 quand $s = 0$ à 1 quand $s = 1$. Chaque fonction génératrice définit un processus de Galton-Watson. Un autre paramètre important est l'espérance du nombre d'enfants,

$$m = \mathbb{E}[Z] = \mathbb{E}[Z_1] = \sum_{k=0}^{\infty} k p_k = f'(1).$$

Intuitivement, on peut penser que si m est supérieur à 1, alors la population explose et si m est inférieur à 1, alors la population disparaît. Nous allons voir que cette intuition est correcte et que la survit de la population ne dépend que de la valeur de m et pas de la distribution. Notons f_n la fonction génératrice de Z_n, Z_n étant la taille de la $n^{\text{ième}}$ génération,

$$f_n(s) = \mathbb{E}[s^{Z_n}],\ 0 \leq s \leq 1.$$

Cette notation permet d'écrire $f_0(s) \equiv s$ et $f_1(s) \equiv f(s)$. Nous voulons maintenant trouver la relation entre f_n et f. Pour cela, nous utilisons des espérances conditionnelles. Si Z_{n-1} est le nombre de particules de la $n-1^{\text{ième}}$ génération, on note $Y_n(1), \ldots, Y_n(Z_{n-1})$ les tailles des descendance de ces particules. L'ensemble de ces tailles représente une suite de variables aléatoires indépendantes et identiquement distribuées, en effet, tous les $Y_n(j)$ ont la même distribution que Z_1 et sont choisis indépendamment. On a alors,

$$\begin{aligned}
f_n(s) &= \mathbb{E}\left[\mathbb{E}[s^{Z_n}|Z_{n-1}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[s^{Y_n(1)+\ldots+Y_n(Z_{n-1})}|Z_{n-1}]\right] \\
&= \mathbb{E}\left[\prod_{j=1}^{Z_{n-1}} \mathbb{E}[s^{Y_n(j)}|Z_{n-1}]\right] \\
&= \mathbb{E}\left[\prod_{j=1}^{Z_{n-1}} \mathbb{E}[s^{Z_1}]\right] \\
&= \mathbb{E}\left[f(s)^{Z_{n-1}}\right] \\
&= f_{n-1}(f(s)) \\
&= \overbrace{f(f(\ldots f}^{n \text{ fois}}(s))).
\end{aligned}$$

On se souvient que f est convexe, croissante et que $f(0) = p_0$ et $f(1) = 1$. Quand $m < 1$, le graphe de $f(s)$ est au dessus de s, donc $f(s) = s$ si et seulement si $s = 1$. Donc pour tout s $f_n(s) \to 1$ et on a $f_n(0) = \mathbb{P}\mathrm{r}[Z_n = 0] \to 1$. Quand $m > 1$, il existe une unique solution $q < 1$ de l'équation $f(s) = s$ (voir la figure 10). Il est alors aisé de constater que pour tout $s \in [0, 1[$, $f_n(s) \to q$ et donc $\mathbb{P}\mathrm{r}[Z_n = 0] \to q$.

Nous nous intéressons maintenant à l'extinction du processus. Si on appelle "extinction" l'événement que tout le processus s'arrête, et pas seulement une génération n particulière, on a,

$$\begin{aligned}
\mathbb{P}\mathrm{r}[\mathit{extinction}] &= \mathbb{P}\mathrm{r}[\text{il existe } n \text{ tel que } Z_n = 0] \\
&= \mathbb{P}\mathrm{r}\left[\cup_{i=1}^{\infty}\{Z_i = 0\}\right] \\
&= \lim_{n\to\infty} \mathbb{P}\mathrm{r}\left[\cup_{i=1}^{n}\{Z_i = 0\}\right] \\
&= \lim_{n\to\infty} \mathbb{P}\mathrm{r}\left[Z_n = 0\right] \\
&= q.
\end{aligned}$$

Ainsi, q est la probabilité d'extinction. Nous pouvons donc énoncer le théorème fondamental des processus de Galton-Watson,

Théorème A.1. *Dans un processus de Galton-Watson, si m est la moyenne de la probabilité de reproduction et f est la fonction génératrice de cette probabilité, on note q la probabilité d'extinction du processus, on a alors,*

i. $q = \lim_{n\to\infty} f^{(n)}(0)$;
ii. $q = 1$ *si et seulement si* $m \leq 1$.

Le processus de branchement est dit sous-critique (*resp.* critique, sur-critique) lorsque $m < 1$ (*resp.* $m = 1$, $m > 1$).

Par ailleurs, le théorème suivant nous donne des indiquations sur la taille des processus de Galton-Watson,

Théorème A.2. *En supposant que $p_1 < 1$, dans un processus de Galton-Watson on a*

$$\mathbb{P}\mathrm{r}[\lim_{n\to\infty} Z_n \in \{0, \infty\}] = 1$$

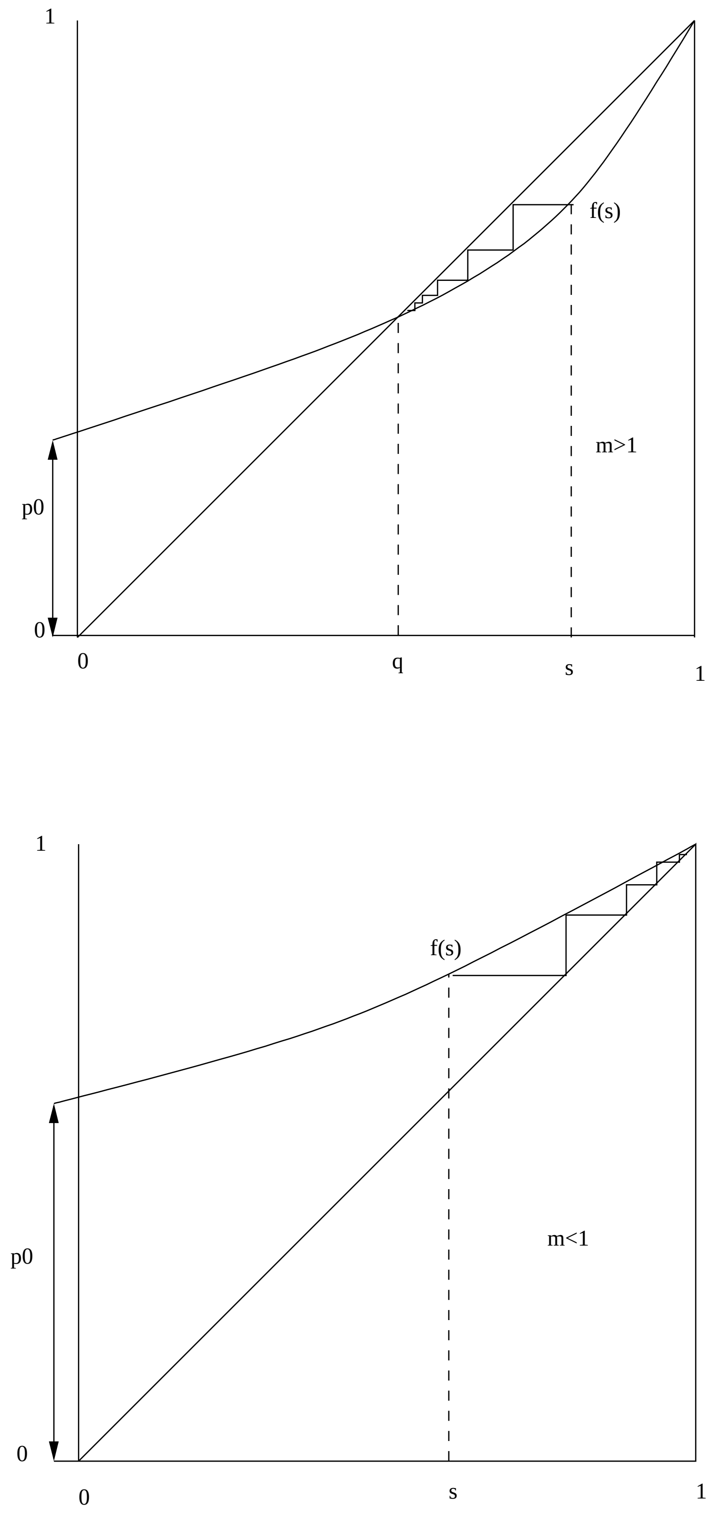

FIGURE 10. Les comportements possibles de la fonction génératrice

Preuve du théorème A.2. On a

$$\mathbb{P}\mathrm{r}[\lim_{n\to\infty} Z_n \notin \{0,\infty\}] \leq \sum_{k=1}^{\infty} \mathbb{P}\mathrm{r}[Z_n = k \text{ infiniment souvent}],$$

cette dernière somme étant nulle si chacun de ses termes est nul, ce qu'il suffit alors de montrer.

Soit r_k la probabilité que la population soit de nouveau de taille k dans le futur alors que nous avons actuellement une population de taille k.

Si $p_0 = 0$, alors

$$r_k \leq \mathbb{P}\mathrm{r}[Z_1 = k | Z_0 = k] = p_1^k < 1.$$

Si $p_0 > 0$, alors

$$r_k \leq \mathbb{P}\mathrm{r}[Z_1 > 0 | Z_0 = k] = 1 - p_0^k < 1.$$

Donc $r_k < 1$.

Soit X le nombre de fois où la population est de taille k. Il faut au moins $n-1$ transitions pour obtenir n populations de taille k à partir d'une population de taille k pour un processus tel que celui considéré, donc

$$\mathbb{P}\mathrm{r}[X \geq n] \leq r_k^{n-1}.$$

On note que

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{P}\mathrm{r}[X \geq n] \leq \sum_{n=0}^{\infty} r_k^n = \frac{1}{1-r_k}.$$

Soit M choisi arbitrairement,

$$\begin{aligned} \mathbb{P}\mathrm{r}[Z_n = k \text{ infiniment souvent}] &\leq \mathbb{P}\mathrm{r}[X \geq M] \\ &\leq \frac{\mathbb{E}[X]}{M} \\ &\leq \frac{1}{M(1-r_k)}, \end{aligned}$$

qui est aussi petit que désiré du fait du choix de M. On conclut donc que

$$\mathbb{P}\mathrm{r}[Z_n = k \text{ infiniment souvent}] = 0.$$

□

3. Décomposition d'un processus de branchement sur-critique

Les résultats que nous présentons maintenant ont été mis sous cette forme par Athreya et Ney [**5**].

Dans le cas sur-critique, la probabilité d'extinction q est strictement inférieure à 1 et q est non nulle si et seulement si $f(0) > 0$. Or beaucoup de résultats sont plus facilement prouvés conditionnellement à la non-extinction du processus, c'est à dire quand $q = 0$ puis, à l'aide d'une transformation due à Harris [**23**] réduit du cas $q = 0$ au cas $q > 0$. Grâce à cette transformation, nous allons pouvoir décomposer le processus de branchement.

Considérons un processus de Galton-Watson sur-critique de fonction génératrice f. Si $f(0) > 0$, on définit la fonction f^* comme suit :

$$f^*(s) = \frac{f((1-q)s+q)-q}{1-q} \text{ pour } 0 \leq s \leq 1.$$

On remarque que $f^*(0) = (f(q)-q)/(1-q) = 0$ et $f^*(1) = 1$. De plus, comme f est une série génératrice à coefficients positifs, f^* l'est aussi et ainsi, f^* est une fonction génératrice. Enfin, on a $f^{*\prime}(s) = f'((1-q)s+q)$ donc $f^{*\prime}(1) = f'(1) = m$.

Soient $(Z_n)_{n\geq 0}$ et $(Z_n^*)_{n\geq 0}$ deux processus de branchement sur le même espace probabiliste $(\Omega, \mathbb{F})$ ayant respectivement comme fonctions génératrices f et f^*.

Soient $\mathbb{P}\mathrm{r}$ et $\mathbb{P}\mathrm{r}^*$ les mesures correspondantes sur $(\Omega, \mathbb{F})$. Si q^* est la probabilité d'extinction de $(Z_n^*)_{n\geq 0}$, on a $q^* = f^*(0) = 0$. Par ailleurs, un simple calcul permet de vérifier que les itérations de la fonction f^* sont :

$$f_n^*(s) = \frac{f_n((1-q)s+q)-q}{1-q} \text{ pour } 0 \leq s \leq 1,\ n \geq 1.$$

On définit les ensembles A et B de la façon suivante :

$$A = \{\omega : Z_n(\omega) \to \infty \text{ quand } n \to \infty\},$$

et

$$B = \{\omega : \exists n \geq 1 \text{ tel que } Z_n(\omega) = 0\}.$$

On a $\mathbb{P}\mathrm{r}[A \cup B] = 1$. D'autre part, $\mathbb{P}\mathrm{r}[B] = q$ par définition, donc, comme par hypothèse $0 < \mathbb{P}\mathrm{r}[B] < 1$, on a $0 < \mathbb{P}\mathrm{r}[A] < 1$. Comme précédemment, pour tout ω, $Z_n(\omega)$ est le nombre de particules appartenant à la $n^{\text{ième}}$ génération de l'arbre correspondant à ω. On définit alors,

$$Z_n^{(1)}(w) = \begin{cases} 0 & \text{si } \omega \in B, \\ \text{le nombre de particules parmi les } Z_n(\omega) \text{ de la } n^{\text{ième}} \text{ génération ayant une descendance infinie,} & \text{si } \omega \in A, \end{cases}$$

et

$$Z_n^{(2)}(\omega) = Z_n(\omega) - Z_n^{(1)}(\omega).$$

3.1. Le sous arbre "infini". En utilisant les définitions posées ci-dessus, on va s'intéresser à l'ensemble A et au processus $Z_n^{(1)}$.

On note $\mathbb{F}_A$ la σ-algèbre sur A obtenue en intersectant les ensembles de $\mathbb{F}$ avec A. Pour tout $E \in \mathbb{F}_A$ on définit $\mathbb{P}\mathrm{r}_A[E] = \mathbb{P}\mathrm{r}[E]/\mathbb{P}\mathrm{r}[A]$. $\mathbb{P}\mathrm{r}_A$ définit alors une mesure de probabilité sur $\mathbb{F}_A$ et ainsi $(A, \mathbb{F}_A, \mathbb{P}\mathrm{r}_A)$ devient un espace probabiliste. On peut alors énoncer le théorème suivant,

Théorème A.3. *Les processus stochastiques $(Z_n^*)_{n\geq 0}$ sur l'espace $(\Omega, \mathbb{F}, \mathbb{P}\mathrm{r}^*)$ et $(Z_n^1)_{n\geq 0}$ sur l'espace $(A, \mathbb{F}_A, \mathbb{P}\mathrm{r}_A)$ sont équivalent au sens des distributions finies dimensionnelles.*

Preuve du théorème A.3. Nous allons montrer le théorème en montrant que le processus $(Z_n^1)_{n\geq 0}$ sur $(A, \mathbb{F}_A, \mathbb{P}\mathrm{r}_A)$ est un processus de Markov ayant les même probabilités de transition que le processus $(Z_n^*)_{n\geq 0}$. Cela suffit car $\mathbb{P}\mathrm{r}_A[Z_0^{(1)} = 1] = 1 = \mathbb{P}\mathrm{r}^*[Z_0^* = 1]$.

Considérons ce qui se passe à la $n^{\text{ième}}$ étape. Sur A, $Z_n^{(1)}(\omega)$ est non nul pour tout n. Chacune des $Z_n^{(1)}(\omega)$ particules va avoir une branche infinie. De plus chacune va avoir un nombre aléatoire de descendants dont au moins un va donner lieu à une branche infinie. Suivant l'une des principales propriétés des processus de branchement, les $Z_n^{(1)}(\omega)$ particules se reproduisent indépendamment les unes des autres et indépendamment du passé et ainsi, $Z_{n+1}^{(1)}(\omega)$, le nombre de particules parmi les $Z_{n+1}(\omega)$ de la $n+1^{\text{ième}}$ génération ayant une descendance infinie, ne dépendent que de la reproduction de $Z_n^{(1)}(\omega)$. Ainsi, le processus $(Z_n^1)_{n\geq 0}$ est markovien, et sachant $Z_n^{(1)} = i$, $Z_{n+1}^{(1)}$ peut être mis sous la forme :

$$Z_{n+1}^{(1)} = \sum_{i=1}^{i} N_j,$$

où pour $1 \leq j \leq i$, N_j est le nombre de particules parmi les descendants de la $j^{\text{ième}}$ particule de $Z_n^{(1)}(\omega)$ ayant une descendance infinie. Comme cela a été indiqué précédemment, les variables aléatoires N_j sont mutuellement indépendantes et donc il faut montrer qu'elles ont la même distribution que Z_1^*. Mais, comme les N_j ont la même distribution que $Z_1^{(1)}$ avec $\mathbb{P}\mathrm{r}_A$, il suffit de montrer que, pour $k \geq 1$,

$$\mathbb{P}\mathrm{r}_A[Z_1^{(1)} = k] = \mathbb{P}\mathrm{r}^*[Z_1^* = k]. \tag{13}$$

Par définition,

$$\mathbb{P}\mathrm{r}_A[Z_1^{(1)} = k] = \frac{\mathbb{P}\mathrm{r}[\{Z_1^{(1)} = k\} \cap A]}{\mathbb{P}\mathrm{r}[A]},$$

où $\mathbb{P}\mathrm{r}$ est la probabilité non conditionnée de départ. Pour $k \geq 1$,

$$\begin{aligned}
\mathbb{P}\mathrm{r}[\{Z_1^{(1)} = k\} \cap A] &= \mathbb{P}\mathrm{r}[Z_1^{(1)} = k] = \sum_{l=k}^{\infty} \mathbb{P}\mathrm{r}[Z_1^{(1)} = k | Z_1 = l]\mathbb{P}\mathrm{r}[Z_1 = l] \\
&= \sum_{l=k}^{\infty} \binom{l}{k} (1-q)^k q^{1-k} \mathbb{P}\mathrm{r}[Z_1 = l].
\end{aligned}$$

Et donc, on a pour tout $s \in [0,1]$,

$$\begin{aligned}
\sum_{l=k}^{\infty} s^k \mathbb{P}\mathrm{r}[Z_1 = l] &= (1-q)^{-1} \sum_{k=1}^{\infty} s^k \sum_{l=k}^{\infty} \binom{l}{k} (1-q)^k q^{1-k} \mathbb{P}\mathrm{r}[Z_1 = l] \\
&= (1-q)^{-1} \sum_{l=1}^{\infty} \mathbb{P}\mathrm{r}[Z_1 = l] \sum_{k=1}^{l} \binom{l}{k} (1-q)^k s^k q^{1-k} \\
&= (1-q)^{-1} \sum_{l=1}^{\infty} \mathbb{P}\mathrm{r}[Z_1 = l](((1-q)s + q)^l - q^l) \\
&= (1-q)^{-1}(f((1-q)s+q) - q) \\
&= f^*(s),
\end{aligned}$$

ce qui permet d'obtenir (13) et achève la démonstration. □

3.2. Le sous arbre fini. Dans la décomposition présentée en 3, $Z_n^{(1)}(\omega)$ ne s'éteint jamais sur A et est identiquement nul sur B. On considère maintenant le comportement de $Z_n^{(2)}(\omega)$ sur B. Ici, $Z_n^{(2)}$ coïncide exactement avec Z_n et on a

Théorème A.4. *Le processus stochastique $Z_n(\omega)$ sur $(B, \mathbb{F}_B, \mathbb{P}\mathrm{r}_B)$ est équivalent à un processus de branchement de Galton-Watson sous-critique $(\hat{Z}_n)_{n \geq 0}$. sur l'espace de probabilité $(\Omega, \mathbb{F}, \mathbb{P}\mathrm{r}^*)$ avec,*

$$\mathbb{P}\mathrm{r}^*[\hat{Z}_0 = 1] = 1 \;\; et \; \mathbb{E}^*[s^{\hat{Z}_1}] = \hat{f}(s) = q^{-1} f(sq).$$

Preuve du théorème A.4. Il est clair que le processus $Z_n(\omega)$ sur $(B, \mathbb{F}_B, \mathbb{P}\mathrm{r}_B)$ est toujours Markovien. On montre que les probabilité de transitions sont les même que $(\hat{Z}_n)$ en vérifiant que

$$\begin{aligned}
\sum_{j=0}^{\infty} \mathbb{P}\mathrm{r}[Z_{n+1} = j | Z_n = i, B] s^j &= \frac{\sum_{j=0}^{\infty} \mathbb{P}\mathrm{r}[Z_{n+1} = j, Z_n = i, B] s^j}{\mathbb{P}\mathrm{r}[Z_n = i, B]} \\
&= \frac{\sum_{j=0}^{\infty} s^j q^j \mathbb{P}\mathrm{r}[Z_{n+1} = j, Z_n = i]}{q^i} \\
&= q^{-i}(f(qs))^i \\
&= (\hat{f}(s))^i.
\end{aligned}$$

Le processus est sous-critique car sa moyenne est

$$\hat{m} = \hat{f}'(1) = f'(q) < 1.$$

Ceci achève la preuve de la démonstration. □

APPENDICE B

La probabilité d'un bicycle

On part de [**10**]

$$q \leq \frac{2m}{n(n-1)} \sum_{s=2}^{n} s^2 \left(\frac{m}{n-1}\right)^s. \tag{14}$$

On pose $c = m/(n-1)$ et donc on cherche à évaluer $\frac{2c}{n} \sum_{s=2}^{n} s^2 c^s$.

On remarque que

$$\sum_{s=2}^{n} s^2 c^s = \sum_{s=2}^{n} (s+1)^2 c^s - 2\sum_{s=2}^{n} sc^s - \sum_{s=2}^{n} c^s. \tag{15}$$

Or,

$$\begin{aligned}\sum_{s=2}^{n} (s+1)^2 c^s &= \frac{1}{c} \sum_{s=1}^{n-1} (s+1)^2 c^{s+1} - 4c + (n+1)^2 c^n \\ &= \frac{1}{c} \sum_{s=2}^{n} s^2 c^s - 4c + (n+1)^2 c^n.\end{aligned}$$

Donc,

$$\left(1 - \frac{1}{c}\right) \sum_{s=2}^{n} s^2 c^s = (n+1)^2 c^n - 4c - 2\sum_{s=2}^{n} sc^s - \sum_{s=2}^{n} c^s. \tag{16}$$

De même on peut calculer $\sum_{s=2}^{n} sc^s$. On a $\sum_{s=2}^{n} sc^s = \sum_{s=2}^{n} (s+1)c^s - \sum_{s=2}^{n} c^s$ et

$$\begin{aligned}\sum_{s=2}^{n} (s+1)c^s &= \frac{1}{c} \sum_{s=1}^{n-1} (s+1)c^s - 2c + (n+1)c^n \\ &= \frac{1}{c} \sum_{s=2}^{n} sc^s - 2c + (n+1)c^n.\end{aligned}$$

D'où on déduit

$$2\sum_{s=2}^{n} sc^s = 2\frac{n+1}{c-1} c^{n+1} - 4\frac{c^2}{c-1} - 2\frac{c}{c-1} \sum_{s=2}^{n} c^s.$$

On remplace alors dans (16) et, avec $\sum_{s=2}^{n} c^s = c^2 \frac{c^{n-1}-1}{c-1}$, on a

$$\begin{aligned}
\left(1-\frac{1}{c}\right)\sum_{s=2}^{n} s^2 c^s &= (n+1)^2 c^n - 4c - 2\frac{n+1}{c-1}c^{n+1} + 4\frac{c^2}{c-1} + \frac{c+1}{c-1}\sum_{s=2}^{n} c^s \\
&= (n+1)^2 c^n - 4c - 2\frac{n+1}{c-1}c^{n+1} + 4\frac{c^2}{c-1} + \frac{c^2(c+1)(c^{n-1}-1)}{(c-1)^2} \\
&= \frac{(n+1)^2(c^{n+1}-c^n) - 4c^2 + 4c - 2(n+1)c^{n+1} + 4c^2}{c-1} + \frac{c^2(c+1)(c^{n-1}-1)}{(c-1)^2} \\
&= \frac{(n^2-1)c^{n+1} - (n+1)^2 c^n + 4c}{c-1} + \frac{c^2(c+1)(c^{n-1}-1)}{(c-1)^2} \\
&= \frac{(n^2-1)c^{n+2} - (n+1)^2 c^{n+1} + 4c^2 - (n^2-1)c^{n+1} + (n+1)^2 c^n - 4c}{(c-1)^2} \\
&\quad + \frac{c^{n+2} + c^{n+1} - c^3 - c^2}{(c-1)^2} \\
&= \frac{1}{(c-1)^2}\left[n^2 c^{n+2} - (2n^2+2n-1)c^{n+1} + (n+1)^2 c^n - c^3 + 3c^2 - 4c\right].
\end{aligned}$$

Donc, en remplaçant dans (15) puis dans (14), on obtient

$$q \leq \frac{2}{n}\frac{c^2}{(c-1)^3}\left[n^2 c^{n+2} - (2n^2+2n-1)c^{n+1} + (n+1)^2 c^n - c^3 + 3c^2 - 4c\right]$$

APPENDICE C

L'amplitude du seuil de transition de 2-SAT aléatoire

1. Introduction

Les résultats qui ont été présentés dans les chapitres 4 et 5 sont des approximations de la largeur de l'intervalle $I(n)$. Bollobàs, Borgs, Chayes, Kim et Wilson [**7**] ont déterminés ultérieurement la taille exacte de $I(n)$ en montrant le théorème suivant,

Théorème C.1 (Bollobàs, Borgs, Chayes, Kim et Wilson, 1999). *Soit $F = F_{n,m}$ une formule 2-CNF aléatoire, conjonction de m clauses différentes, sur n variables, si on pose $m = (1+\varepsilon)n$ et $\varepsilon = \lambda_n n^{-1/3}$. Alors il existe des constantes ε_0 et λ_0 et $0 < \varepsilon_0 < 1$, $0 < \lambda_0 < \infty$ telles que, dès que $|\lambda_n| \leq \varepsilon_0 n^{1/3}$,*

$$\mathbb{P}\mathrm{r}[F_{n,m} \textit{ est satisfaisable}] = \begin{cases} 1 - \Theta\left(\frac{1}{|\lambda_n^3|}\right) & \textit{si } \lambda_n < -\lambda_0, \\ \Theta(1) & \textit{si } |\lambda_n| \leq \lambda_0, \\ \exp(-\Theta(|\lambda_n|^{-3})) & \textit{si } \lambda_n > \lambda_0. \end{cases}$$

Ce théorème implique que la probabilité de satisfaisabilité ainsi que la probabilité d'insatisfaisabilité de $F_{n,m}$ sont supérieures à δ quand le rapport m/n appartient à l'intervalle $[1 - C_\delta n^{-1/3}, 1 + C_\delta n^{-1/3}]$, C_δ étant une constante positive spécifiée. Ce résultat est le meilleur que l'on puisse espérer car il n'est plus vrai lorsqu'on remplace C_δ par une constante positive suffisamment grande.

Nous allons maintenant exposer un schéma de la preuve de ce théorème. À cette fin, nous allons utiliser le graphe dirigé utilisé précédemment et introduire diverses structures liées à ce graphe. La présence de ces structures est liées à la satisfaisabilité de la formule et leur probabilité d'existence permet de déterminer la probabilité de satisfaisabilité de la formule.

2. L'épine et le graphe allégé

Soit une formule F sous forme $2CNF$. Comme précédemment nous allons considérer la représentation de cette formule sous forme de graphe dirigé comme cela a été présenté dans la section 5 du chapitre 1. Dans la suite nous noterons D_F le graphe associé à la formule F.

2.1. L'épine d'une formule. La définition de l'épine d'une formule a été donnée dans la section 2 du chapitre 2. Il nous reste cependant à montrer le lemme 2.6. On rappelle que ce lemme établit que l'épine d'une formule F correspond, dans le graphe associé à F, à l'ensemble des sommets u tels qu'il existe une arête entre u et $\neg u$.

Preuve du lemme 2.6. Supposons que $u \rightsquigarrow \neg u$ et soit $u = v_0 \to v_1 \to \ldots \to v_{r-1} \to v_r = \neg u$ le chemin le plus court de u à $\neg u$ dans D_F. Aucun littéral n'apparaît deux fois dans ce chemin, bien que l'on puisse avoir un littéral et sa négation. Soit l le plus petit entier positif tel que v_l n'est strictement distinct de tous les littéraux $v_0, \ldots, v_{l-1}$ et soit $k \in \{0, l-1\}$ tel que $v_l = \neg v_k$. Alors

$H = (\neg v_0 \vee v_1) \wedge (\neg v_1 \vee v_2) \wedge \ldots \wedge (\neg v_{l-1} \vee v_l)$ est une sous-formule de F qui est satisfaite en assignant **faux** à chacun des littéraux $v_0, v_1, \ldots, v_{l-1}$. D'autre part, $H \wedge u = H \wedge v_0$ n'est pas satisfaisable puisque afin de satisfaire cette formule, il faudrait assigner **vrai** à v_0, puis **vrai** à v_1 et ainsi de suite jusqu'à assigner **vrai** à v_l. Toutefois, comme v_k est **vrai**, v_l est déjà **faux**. Nous avons donc montré que $u \rightsquigarrow \neg u$ implique le fait que $u \in S(F)$.

Inversement, on suppose que $H \subset F$ est satisfaisable $H' = H \wedge u = H \wedge (u \vee u)$ est insatisfaisable. Alors $D_{H'}$ contient un cycle contradictoire $C = u \rightsquigarrow \neg u \rightsquigarrow u$. Par ailleurs D_H ne contient pas de cycle contradictoire, le cycle C de H' contient l'arc $\neg u \to u$, c'est à dire $v \rightsquigarrow \neg v \rightsquigarrow \neg u \to u \rightsquigarrow v$. Alors, nous avons $u \rightsquigarrow v \rightsquigarrow \neg v \rightsquigarrow \neg u$ dans D_H, donc $u \rightsquigarrow \neg u$. Ainsi, si $u \in S(F)$, alors $u \overset{D_F}{\rightsquigarrow} \neg u$. □

Le lemme suivant donne une autre interprétation de l'épine d'une formule sous forme $2CNF$. Nous avons besoin pour cela d'introduire le graphe *sortant* $D_F^+(u)$ d'un sommet u dans D_F, qui est l'ensemble des sommets et arêtes accessibles à partir de u. L'ensemble des sommets de $D_F^+(u)$ est

$$L^+(u) = L_F^+(u) = \{v | u \overset{D_F}{\rightsquigarrow} v\},$$

et $D_F^+(u)$ contient tous les arcs $v \to w$ de D_F tels que $v \in L_F^+(u)$. De même, on définit le graphe *entrant* $D_F^-(u)$ d'un sommet u et l'ensemble des sommets de ce graphe est

$$L^-(u) = L_F^-(u) = \{v | v \overset{D_F}{\rightsquigarrow} u\},$$

Dans la suite nous dirons qu'un ensemble est *strictement distinct* si les littéraux de cet ensemble sont strictement distincts deux à deux.

Lemme C.2. *Pour toute formule sous forme $2CNF$,*

$$\begin{aligned} S(F) &= \{u | L_F^+(u) \text{ n'est pas strictement distinct}\} \\ &= \{u | L_F^+(u) \backslash \{u, \neg u\} \text{ n'est pas strictement distinct}\}. \end{aligned}$$

Preuve du lemme C.2. On commence par montrer la première égalité. On se rappellera tout d'abord que, par le lemme 2.6 on a $S(F) = \{u | u \overset{D_F}{\rightsquigarrow} \neg u\}$. Soit un littéral u tel que $u \rightsquigarrow \neg u$, alors $\{u, \neg u\} \subset L_F^+(u)$ donc $L_F^+(u)$ n'est pas strictement distinct. Inversement, si $L_F^+(u)$ n'est pas strictement distinct, il existe un littéral v tel que $\{v, \neg v\} \subset L_F^+(u)$ et ainsi $u \rightsquigarrow v$ et $u \rightsquigarrow \neg v$. D'après la construction du graphe associé à F, ceci implique $v \rightsquigarrow \neg u$ et comme $u \rightsquigarrow v$ on a finalement $u \rightsquigarrow \neg u$. Ceci montre la première égalité.

Avant de montrer la seconde égalité, on note tout d'abord que l'ensemble des littéraux u pour lesquels $L_F^+(u) \backslash \{u, \neg u\}$ n'est pas strictement distinct est un sous ensemble de l'ensemble des littéraux u pour lesquels $L_F^+(u)$ n'est pas strictement distinct. Il nous reste alors à montrer que si $L_F^+(u)$ n'est pas strictement distinct, alors $L_F^+(u) \backslash \{u, \neg u\}$ n'est pas strictement distinct. Supposons donc que $L_F^+(u)$ n'est pas strictement distinct. La première égalité du lemme nous permet d'écrire $u \rightsquigarrow \neg u$. Comme le graphe D_F ne contient pas l'arc allant de u à $\neg u$ (ceci impliquerai la présence de la clause $u \wedge u$ dans la formule), il existe un littéral v strictement distinct de u tel que $u \rightsquigarrow v \rightsquigarrow \neg u$. Cela implique $u \rightsquigarrow v$ et $u \rightsquigarrow \neg v$, donc $L_F^+(u) \backslash \{u, \neg u\}$ n'est pas strictement distinct. □

On remarque que la seconde partie de la preuve suppose l'absence de clauses de taille un. La seconde égalité du lemme C.2 n'est donc pas valable pour des formules contenant des clauses de taille un et des clauses de taille deux, contrairement à la première égalité.

2.2. Le graphe allégé. Nous allons maintenant présenter la construction d'une version allégée du graphe sortant $D_F^+(u)$. Nous noterons $\tilde{D}_F^+(u)$ ce nouveau graphe et $\tilde{L}_F^+(u)$ son ensemble de sommets. Ce graphe est important car il existe une relation entre celui-ci et un graphe aléatoire non orienté et dont les arêtes sont indépendantes qui sera précisé plus tard.

Ce graphe sortant allégé $\tilde{D}_F^+(u)$ est construit en effectuant un recherche locale dans D_F à partir du littéral u. Le *graphe courant* sera le sous graphe de D_F qui a déjà été examinés par la recherche locale. La *frontière* est constituée des sommets du graphe courant pour lesquels l'exploration n'est pas terminée. Initialement, le graphe courant est constitué du littéral u et u est dans la frontière. Lorsque la frontière est vide, la recherche locale prend fin et $\tilde{D}_F^+(u)$ est le graphe courant. La recherche s'effectue comme suit,

1. un littéral v est choisi arbitrairement dans la frontière ;
2. pour chaque littéral w tel que ni w, ni $\neg w$ n'est dans le graphe courant, on vérifie si $v \to w$ est dans D_F ou pas ;
3. pour chaque littéral w tel que $v \to w$ est dans D_F, on étiquette w comme un "nouveau littéral", à moins que cela soit le cas pour w et $\neg w$, auquel cas un seul des deux sera un "nouveau littéral" (celui qui n'est pas nié par exemple) ;
4. on ajoute chaque nouveau littéral w et l'arc $v \to w$ au graphe courant ;
5. on ajoute chaque nouveau littéral w à la frontière. On enlève v de la frontière ;
6. on considère la paire de sommets ordonnée (w, f) telle que soit w est nouveau et f est dans la frontière mais pas nouveau, soit w et f sont tous les deux nouveaux et w est lexicographiquement plus petit que f. On teste si les arcs $w \to f$ ou $f \to w$ sont dans D_F. Si l'un d'eux est présent, on l'ajoute au graphe courant et si les deux sont présents on en ajoute seulement un dans le graphe courant.

Nous présentons maintenant le graphe aléatoire non orienté associé au graphe allégé. Le modèle de formule utilisé est le modèle $F_{n,p}$ (voir la section 7 du chapitre 1). En fait nous allons comparer la distribution de $\tilde{D}_F^+(u)$ avec celle de composante connexe d'un certain sommet du graphe aléatoire $G_{n,2p-p^2}$, où de façon classique $G_{n,\tilde{p}}$ est le graphe aléatoire dont l'ensemble des sommets est $\{x_1, \dots, x_n\}$ et qui est obtenu à partir du graphe complet sur $\{x_1, \dots, x_n\}$ en gardant chaque sommet avec probabilité $\tilde{p}$. À cette fin, nous allons transformer le graphe orienté $\tilde{D}_F^+(u)$ en un graphe non orienté de sommets $x_1, \dots, x_n$. Tout d'abord, pour $i \in \{1, \dots, n\}$ nous associons les littéraux u_i et $\neg u_i$ avec la variable correspondante x_i. Ensuite, chaque arc $u_i \to u_j$, $i, j \in \{1, \dots, n\}$, $i \neq j$, donne lieu à une arête non orientée $\{x_i, x_j\}$. Le graphe obtenu sera noté $\tilde{D}_F$. Nous noterons $\langle x \rangle$ la variable associée au littéral x et $C_{n,2p-p^2}(x)$ sera la composante connexe contenant le sommet $\langle x \rangle$ dans le graphe $G_{n,2p-p^2}$.

Nous pouvons maintenant énoncer le lemme C.3,

Lemme C.3. *Le graphe allégé $\tilde{D}_F^+(u)$ défini ci-dessus a les propriété suivantes :*

i. *$\tilde{D}_F^+(u)$ est un sous-graphe de $D_F^+(u)$;*
ii. *$\tilde{L}_F^+(u)$ est strictement distinct ;*
iii. *$\tilde{L}_F^+(u) = L_F^+(u)$ si et seulement si $L_F^+(u)$ est strictement distinct ;*
iv. *pour une formule $F = F_{n,p}$, le graphe $\tilde{D}_F(u)$ a la même distribution que $C_{n,2p-p^2}(u)$. En particulier, $|\tilde{L}_{F_{n,p}}^+(u)|$ et $|C_{n,2p-p^2}(u)|$ sont equidistribués.*

Preuve du lemme C.3. Les propriétés i et ii sont vraies par construction.

Pour la propriété iii, on suppose tout d'abord que $L_F^+(u)$ est strictement distinct. On remarque qu'il existe certains arcs permettant de sortir de l'ensemble de sommets $\tilde{L}_F^+(u)$ qui ne sont pas testés, ou bien qui sont testés et présents mais qui néanmoins n'apparaissent pas dans le graphe allégé $\tilde{D}_F^+(u)$. Chacun de ces arcs mène à un littéral déjà dans $\tilde{L}_F^+(u)$, ou dont la négation est déjà dans $\tilde{L}_F^+(u)$. Ainsi, si l'ensemble de littéraux $L_F^+(u)$ contenait plus de littéraux que $\tilde{L}_F^+(u)$, ceci seraient des négations de littéraux déjà présents et $L_F^+(u)$ ne serait pas strictement distinct. Inversement, si $\tilde{L}_F^+(u)$ et $L_F^+(u)$ sont égaux, $L_F^+(u)$ est strictement distinct par la propriété ii.

Prouvons enfin la propriété iv. Pour chaque littéral $u \in \{1, \ldots, n\}$, on définit $[u] = \{u, \neg u\}$. Nous allons montrer par récurrence sur l'exécution de l'algorithme que les propriétés suivantes sont vraies au début et à la fin de chaque pas d'exécution,

1. pour chaque paire de littéraux u et v du graphe courant, exactement deux arcs entre $[u]$ et $[v]$ ont été testés ;
2. pour tout littéral v appartenant au graphe courant mais pas à la frontière, et tout littéral w tel que ni w ni $\neg w$ n'appartient au graphe courant, exactement deux arcs entre $[v]$ et $[w]$ ont été testés, chacun avec un résultat négatif ;
3. pour chaque littéral v dans la frontière, et tout littéral w tel que ni w ni $\neg w$ n'est dans le graphe courant, aucun des arcs entre $[v]$ et $[w]$ n'a été testé ;
4. si ni u,ni $\neg u$, ni v et ni $\neg v$ ne font partie du graphe courant, alors aucun arc entre $[u]$ et $[v]$ n'a été testé ;
5. pour toute paire de littéraux strictement distincts (u, v), soit aucun, soit exactement un des quatre arcs entre $[u]$ et $[v]$ n'apparaît dans le graphe courant. Un arc apparaît si et seulement si un test entre $[u]$ et $[v]$ a rendu un résultat positif.

Supposons que les conditions 1 à 5 sont vérifiées au début d'un pas d'exécution. Pour vérifier que 1 est toujours valable à la fin du pas, on remarque que aucun arc entre $[u]$ et $[v]$ n'a été testé dans le pas actuel si ni u ni v n'est nouveau. Si v est vieux et u est nouveau, alors soit v était le sommet sélectionné dans la frontière, auquel cas les arcs $v \to u$ et $v \to \neg u$ ont été testés dans le pas actuel, soit v n'était pas dans la frontière et les arcs entre $[u]$ et $[v]$ ont été testés durant un pas précédent par l'hypothèse 2. Si u et v sont tous les deux nouveaux, alors par l'hypothèse 3, aucun arc entre $[u]$ et $[v]$ n'a été testé durant un pas précédent et seulement les deux arcs $u \to v$ et $v \to u$ sont testés durant le pas actuel.

Pour montrer 2, on observe que si v est dans le graphe courant mais pas dans la frontière, il était dans la frontière durant un pas précédent et en a été ôté après que tous les arcs de v vers les sommets u tels que l'on ait pas u et $\neg u$ simultanément dans le graphe courant n'aient été testés. cela inclus en particulier le sommet w en question, et puisque l'on suppose que ni w ni $\neg w$ ne sont dans le graphe courant, il s'ensuit que les deux tests ont donnés des résultats négatifs. Après ce pas, v n'est plus dans la frontière donc aucune arête contenant v ou $\neg v$ n'est plus jamais testée, ce qui implique la propriété 2.

L'observation du fait qu'un arc entre un sommet v du graphe courant et un sommet w, tel que ni w ni $\neg w$ n'est dans le graphe courant, n'est testé que si v est le sommet choisi durant le pas actuel et n'est donc plus dans la frontière ultérieurement donne la propriété 3.

La propriété 4 suit simplement du fait que, puisqu'un arc $f \to v$ est testé seulement si f est dans la frontière (et donc dans le graphe courant au pas actuel), ou si f et v sont de nouveaux sommets, ce qui signifie qu'ils sont dans le graphe courant après une recherche locale.

Pour montrer que la propriété 5 est vraie à la fin du pas, on considère trois cas. Dans le premier cas, aucun des sommets u, $\neg u$, v ou $\neg v$ n'est dans le graphe courant, auquel cas aucune arête entre $[u]$ et $[v]$ n'apparaît dans le graphe courant par l'hypothèse 4. Le second cas est lorsque l'un exactement des quatre littéraux u, $\neg u$, v ou $\neg v$ est dans le graphe courant. Supposons que c'est le cas de v. Alors, par les hypothèses 2 et 3 aucun des arcs entre $[u]$ et $[v]$ n'apparaît dans le graphe courant. Le troisième cas est lorsque deux des sommets u, $\neg u$, v ou $\neg v$ sont dans le graphe courant, par exemple u et v. Alors par l'hypothèse 1, deux des arcs entre $[u]$ et $[v]$ ont été testés et comme la procédure teste toujours deux arêtes en même temps et en ajoute une seule, on obtient 4.

Les propriétés 1 à 5 permettent alors de montrer la partie iv du lemme. Si on tire toutes les paires non ordonnées dans un ordre arbitraire en attribuant à chaque coup "présent" avec probabilité $2p - p^2$ ou absent avec probabilité $(1-p)^2$, le résultat sera le graphe aléatoire $G_{n,2p-p^2}$. C'est ce que fait la recherche locale, en s'arrêtant lorsque la composante connexe contenant $\langle x \rangle$ a été déterminée. Ainsi si on considère $\tilde{D}_F^+(u)$ comme un graphe non orienté, c'est la composante connexe contenant $\langle x \rangle$ dans $G_{n,2p-p^2}$. □

3. La probabilité de satisfaisabilité

On considère ici une formule F choisie aléatoirement selon le modèle $F_{n,p}$ et nous allons montrer le théorème suivant,

Théorème C.4. *Il existe des constantes λ_0 et ε_0, $0 < \lambda_0 < \infty$ et $0 < \varepsilon_0 < 1$, tel que pour $p = \frac{1}{2n}(1 + \lambda_n n^{-1/3})$ et $\lambda_0 \leq |\lambda_n| \leq \varepsilon_0 n^{1/3}$ on ait,*

i. si $\lambda_n < 0$, alors

$$\mathbb{P}\mathrm{r}[F_{n,p} \text{ est satisfaisable}] = \exp(-\Theta(|\lambda_n|^{-3}));$$

ii. si $\lambda_n > 0$, alors

$$\mathbb{P}\mathrm{r}[F_{n,p} \text{ est satisfaisable}] = \exp(-\Theta(|\lambda_n|^{3})).$$

Ce théorème et la section 7 du chapitre 1 permettent l'obtention du théorème C.1.

Afin de montrer le théorème C.4, nous devons considérer la formule aléatoire $F_{n,p}$ comme un processus (on construit la formule clause par clause) et coupler les processus pour toutes les valeurs de n et p. À cette fin, pour chaque couple d'entiers naturels on tire quatre variables aléatoires distribuées uniformément dans $[0,1]$. On note $\binom{X}{2}$ l'ensemble de toutes les paires non ordonnées d'éléments de X. Ainsi, l'ensemble de toutes ces variables aléatoires, indexées par 4 copies de $\binom{\mathbb{N}}{2}$ est un ensemble de variables aléatoires indépendantes. Une paire d'entiers spécifie deux variables différentes de la formule, et les quatre variables aléatoires associées à la paire correspondent aux quatre clauses différentes que l'on peut construire avec ces variables de telle sorte que chaque clause C ait sa propre variable aléatoire U_C comprise entre 0 et 1. Une clause apparaît dans une formule sur n variables avec la probabilité p quand les valeurs des indices de ses deux variables ne sont pas supérieures à n et que $U_C < p$. Nous appellerons U_C l'*anniversaire* de la clause C. Le processus $\tilde{F}$ est la collection de toutes ces variables aléatoires et définit une famille de formules de la façon suivante

$$F_{n,p} = \bigwedge_{C : n(C) \leq n, U_C < p} C,$$

où $n(C)$ représente le maximum des valeurs des indices des deux variables de la clause C. On remarque que pour chaque valeur de n et p, la distribution de la formule $F_{n,p}$ est exactement la distribution présentée précédemment, ce qui justifie l'emploi de la même notation. Par construction, la satisfaisabilité de $F_{n,p}$ est monotone, décroissante en n et p.

3.1. Résultats intermédiaires. Avant de montrer le théorème C.4 nous allons énoncer deux lemmes dont les résultats seront utiles. Les démonstrations ne seront pas présentées ici du fait de leur longueur et de leur aspect excessivement techniques, le lecteur intéressé peut néanmoins consulter l'article original de Bollobás *et al.* [**7**]. Soit $\mathcal{V} : \mathbb{R}_+^* \to]0,1[$ la fonction telle que

$$1 - \mathcal{V}(\varepsilon) = \exp[-(1+\varepsilon)\mathcal{V}(\varepsilon)],$$

où encore,

$$\mathcal{V}(\varepsilon) = 1 - \sum_{k=0}^{\infty} \frac{k^{k-1}}{k!}[(1+\varepsilon)e^{-(1+\varepsilon)}]^k.$$

On peut remarquer que pour ε suffisamment petit, on a $\mathcal{V}(\varepsilon) = 2\varepsilon + O(\varepsilon^2)$.

Lemme C.5. *Il existe des constantes λ_0 et ε_0, $0 < \lambda_0 < \infty$ et $0 < \varepsilon_0 < 1$, telles que si $p = \frac{1}{2n}(1 + \lambda_n n^{-1/3})$ et $\lambda_0 \leq |\lambda_n| \leq \varepsilon_0 n^{1/3}$,*

i. si $\varepsilon < 0$, alors

$$\Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x\right] = \frac{n^{-1/3}}{4\lambda_n^2}(1 + o_{\varepsilon,\lambda_n}(1)),$$

ii. si $\varepsilon > 0$, alors

$$\begin{aligned}\Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x\right] &= \mathcal{V}(\varepsilon)(1 + o_{\lambda_n}(1)) \\ &= 2\lambda_n n^{-1/3}(1 + o_{\lambda_n}(1) + O(\varepsilon)).\end{aligned}$$

Les égalités présentées dans ce lemme permettent d'obtenir une estimation de l'espérance de la taille l'épine tandis que le résultat du lemme suivant nous donne une estimation de la variance la taille de l'épine.

Lemme C.6. *Soient p,ε et λ_n comme dans le lemme C.5. Alors, pour tous les littéraux x et y strictement distincts,*

i. si $\varepsilon < 0$, alors

$$\Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x\right]\Pr\left[y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] \leq \Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x \textit{ et } y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] = O\left(\frac{n^{-2/3}}{\lambda_n^4}\right), \tag{17}$$

ii. si $\varepsilon > 0$, alors

$$0 \leq \Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x \textit{ et } y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] - \Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x\right]\Pr\left[y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] = O\left(\frac{n^{-2/3}}{\lambda_n}\right).$$

On remarque que par monotonicité la borne (17) peut être étendue à tous les $\lambda_n \in [-n^{-1/3}, -\lambda_0]$. En effet, les évènements $x \rightsquigarrow \neg x$ et $y \rightsquigarrow \neg y$ étant monotones, on a

$$\Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x \text{ et } y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] \leq \Pr\left[x \overset{F_{n,p_0}}{\rightsquigarrow} \neg x \text{ et } y \overset{F_{n,p_0}}{\rightsquigarrow} \neg y\right]$$

pour tout $p \leq p_0$. Si on pose $p_0 = (1-\varepsilon_0)/2n$, l'équation (17) permet de borner le membre de droite par $O(n^{-2/3}/(n^{1/3})^4)$. Or, $n^{-2/3}/(n^{1/3})^4 = O(n^{-2/3}/\lambda_n)$ pour $\lambda_n \in [-n^{-1/3}, -\varepsilon_0 n^{1/3}]$, on obtient alors

$$\Pr\left[x \overset{F_{n,p}}{\rightsquigarrow} \neg x \text{ et } y \overset{F_{n,p}}{\rightsquigarrow} \neg y\right] = O\left(\frac{n^{-2/3}}{\lambda_n^4}\right) \text{ pour tout } \lambda_n \in [n^{-1/3}, -\lambda_0].$$

3.2. Les bornes inférieures. Afin d'obtenir les bornes inférieures du théorème C.1, il suffit de considérer le couplage précédent pour n fixé. On ne fait donc plus référence à n. Étant donné $\tilde{F}$ (c'est à dire F_p pour chaque p), on définit la formule réduite $\tilde{\Phi} = (\Phi_p)_{p\in[0,1]}$ comme suit : Φ_0 n'a pas de clauses ; on commence avec $p = 0$ et on l'augmente jusqu'à $p = 1$, on ajoute les clauses à F_p une par une et à chaque ajout d'une clause, on ajoute aussi celle-ci à Φ_p à condition que Φ_p e devienne pas insatisfaisable.

Étant donné $\tilde{\Phi}$, nous pouvons définir un nouveau processus $\tilde{H}$ comme suit : pour chaque clause C apparaissant dans $\tilde{\Phi}$, on pose que son anniversaire dans $\tilde{H}$ coïncide avec celui dans $\tilde{\Phi}$. Pour chaque clause C qui n'apparaît pas dans Φ_1, il existe une plus petite valeur $p_{\min}$ de p pour laquelle $\Phi_p \wedge C$ n'est pas satisfaisable. On choisit l'anniversaire de C dans H_p uniformément dans l'intervalle $[p_{\min}, 1[$. Puisque le processus $\tilde{H}$ est tiré uniformément aléatoirement dans l'ensemble des processus $\tilde{F}$ avec la formule réduite $\tilde{\Phi}$, les processus $\tilde{H}$ et $\tilde{F}$ sont identiquement distribués. On remarque que la première fois que H_p diffère de Φ_p est aussi la première fois que H_p devient insatisfaisable.

On note U_F le nombre de clauses C différentes telles que $F \wedge C$ soit insatisfaisable. Si F est satisfaisable, alors

$$u(F) = \binom{|S(F)|}{2},$$

où $S(F)$ est l'épine de F.

Conditionnellement au fait que le processus réduit soit $\tilde{\Phi}$ et à la satisfaisabilité du processus H_p (c'est à dire $H_p = \Phi_p$), la probabilité que $H_{p+\delta}$ soit satisfaisable est,

$$\left(1 - \frac{\delta}{1-p}\right)^{U(\Phi_p)} = \exp\left[-\frac{U(\Phi_p)}{1-p}\delta + O(\delta^2)\right],$$

à condition que δ soit suffisamment petit. Si on multiplie ces probabilité pour $p = 0, \delta, 2\delta, \ldots$ et en passant à la limite $\delta \to 0$, on trouve que conditionnellement à $\tilde{\Phi}$, la probabilité que H_p soit satisfaisable est,

$$\mathbb{P}\mathrm{r}[H_p \text{ est satisfaisable } |\tilde{\Phi}] = \exp\left[-\int_0^p \frac{U(\Phi_p)}{1-s}ds\right].$$

On a alors,

$$\begin{aligned}
\mathbb{P}\mathrm{r}[F_p \text{ est satisfaisable}] &= \mathbb{P}\mathrm{r}[H_p \text{ est satisfaisable}] \\
&= \mathbb{E}_{\tilde{\Phi}}[\mathbb{P}\mathrm{r}[H_p \text{ est satisfaisable } |\tilde{\Phi}]] \\
&= \mathbb{E}_{\tilde{\Phi}}\left[\exp\left[-\int_0^p \frac{U(\Phi_p)}{1-s}ds\right]\right] \\
&\geq \exp \mathbb{E}_{\tilde{\Phi}}\left[-\int_0^p \frac{U(\Phi_p)}{1-s}ds\right] \\
&\geq \exp\left[-\int_0^p \frac{\mathbb{E}_{\tilde{\Phi}}[U(\Phi_p)]}{1-s}ds\right] \\
&\geq \exp\left[-\int_0^p \frac{4\binom{n}{2}\mathbb{P}\mathrm{r}[x \rightsquigarrow \neg x \text{ et } y \rightsquigarrow \neg y \text{ dans } \Phi_s]}{1-s}ds\right] \\
&\geq \exp\left[-4\binom{n}{2}\int_0^p \frac{\mathbb{P}\mathrm{r}[x \rightsquigarrow \neg x \text{ et } y \rightsquigarrow \neg y \text{ dans } F_s]}{1-s}ds\right],
\end{aligned}$$

où x et y sont strictement distincts.

Nous allons maintenant calculer l'intégrale. Nous sommes particulièrement intéressés par le cas $p = O(1/n)$, aussi nous supposons $p = o_n(1)$ de telle manière que l'effet du dénominateur soit négligeable. Par hypothèse du théorème C.4 on a $p = \frac{1}{2n}(1 + \lambda_n n^{-1/3})$ et si on pose

$$s = s(t) = \frac{1 + tn^{-1/3}}{2n},$$

on obtient

(18) $\mathbb{P}\mathrm{r}[F_p$ est satisfaisable$] \geq$

$$\exp\left[-(1 + o_n(1))n^{2/3} \int_{-n^{1/3}}^{\lambda_n} \mathbb{P}\mathrm{r}[x \rightsquigarrow \neg x \text{ et } y \rightsquigarrow \neg y \text{ dans } F_s(t)]dt\right].$$

Preuve de la borne inférieure du théorème C.4 en régime sous-critique. Si $t \in [-n^{1/3}, \lambda_0]$, la probabilité dans l'intégrale de l'équation (18) est, d'après le lemme C.6 et la remarque le suivant, $O(n^{-2/3}/t^4)$. En remplaçant et en intégrant on trouve,

$$\mathbb{P}\mathrm{r}[F_p \text{ est satisfaisable }] \geq \exp\left[\Theta\left(\frac{1}{\lambda_n^3}\right)\right] = 1 - \Theta\left(\frac{1}{|\lambda_n|^3}\right)$$

pour $\lambda_n \in [-n^{-1/3}, -\lambda_0]$. □

Preuve de la borne inférieure du théorème C.4 en régime sur-critique. En utilisant les lemmes C.5 et C.6 on peut affirmer que la valeur de la probabilité dans l'intégrale de l'équation (18) est $4n^{-2/3}t^2(1+O(\varepsilon)+o_{\lambda_n}(1))$, quand $t \in [\lambda_0, \varepsilon_0 n^{1/3}]$. Si $t \in [-\lambda_0, \lambda_0]$ on majore la probabilité par $\Theta(n^{-2/3})$, et pour $t \in [-n^{1/3}, -\lambda_0]$, la probabilité est majorée par $O(n^{-2/3}/t^4)$. En intégrant, on trouve alors

$$\mathbb{P}\mathrm{r}[F_p \text{ est satisfaisable }] \geq \exp\left[-\frac{4 + o_{\varepsilon,\lambda_n}(1)}{3}\lambda_n^3\right]$$

pour $\lambda_n \in [\lambda_0, \varepsilon_0 n^{1/3}]$. □

3.3. Les bornes supérieures. Pour déterminer les bornes supérieures de la probabilité de satisfaisabilité nous allons utiliser le théorème C.8 suivant. Ce théorème établit qu'il existe certaines structures contenues dans le graphe dirigé associé à la formule, avec grande probabilité.

Définition C.7. *Un* hourglass *est un triplet (v, I_v, O_v) où v est un littéral et I_v et O_v sont deux ensembles de littéraux ne contenant pas v tels que pour tout $x \in I_v$, il existe un chemin $x \rightsquigarrow y$ dans $I_v \cup \{v\}$, et pour tout $y \in O_v$, il existe un chemin $v \rightsquigarrow y$ dans $O_v \cup \{v\}$. De plus, $\{v\} \cup I_v \cup O_v$ est strictement distinct. v est le* littéral central, *I_v est la* portion d'entrée *et O_v est la* portion de sortie *du hourglass.*

On remarque que les hourglass correspondent aux graphes entrant et sortant dont il a été question dans la section 2 du chapitre 5. La preuve qu'il en existe de suffisamment grands pour effectuer la preuve ci-dessous n'est pas présentée ici. Toutefois il s'agit aussi de les construire pas à pas en ayant réservé une quantité de sommets pour conserver des probabilités constantes puis en modifiant une preuve de Karp [**26**] de montrer qu'il existe un sommet dont les graphes entrant et sortant sont suffisamment grands avec une probabilité suffisante. La transformation de la preuve de Karp consiste à augmenter cette probabilité afin de pouvoir montrer le théorème C.4.

Théorème C.8. *Il existe des constantes λ_0 et ε_0, $0 < \lambda_0 < \infty$ et $0 < \varepsilon_0 < 1$ tels que pour $p = \frac{1}{2n}(1 + \lambda_n n^{-1/3})$ et $\lambda_0 \leq |\lambda_n| \leq \varepsilon_0 n^{1/3}$ alors avec probabilité au moins égale à $1 - exp(-\Theta(|\lambda_n|^3))$,* *i. si* $\lambda_n < 0$*, alors il y a au moins* $\Theta(|\lambda_n|^3))$ *hourglass disjoints et strictement distincts tels que les portions d'entrée et de sortie soient de taille au moins* $n^{2/3}/\lambda_n$ *;*

ii. si $\lambda_n > 0$*, alors il existe au moins un hourglass dont les portions d'entrée et de sortie sont chacune de taille* $\Theta(\lambda_n n^{2/3})$*.*

La preuve de ce théorème peut aussi être trouvée dans l'article de Bollobás *et al.* [**7**] Ici, les hourglass vont nous permettre de déterminer les bornes supérieures de satisfaisabilité des deux cotés du seuil.

Preuve de la borne supérieure du théorème C.4 en régime sous-critique. Dans la suite nous augmentons p de $(1 - tn^{-1/3})/2n$ à $(1 - (t/2)n^{-1/3}/2n$, avec $\lambda_0 \leq t \leq \varepsilon_0 n^{1/3}$. Pour toute paire de sommets, qu'il existe où non une arête entre eux avant, la probabilité qu'il y en ait une à la fin de la construction est au moins $(t/4)n^{-4/3}$. Pour chaque hourglass, pour chaque paire de littéraux u et v dans la portion de sortie du hourglass, si la clause $\neg u \wedge \neg v$ apparaît, alors le sommet central et toute la portion d'entrée du hourglass font partie de l'épine de la formule à la fin du processus. En effet, soit x un tel sommet, alors $x \rightsquigarrow u$ et $x \rightsquigarrow v$ puisque u et v sont dans la portion de sortie du hourglass. Mais l'apparition de la clause $\neg u \wedge \neg v$ implique $u \rightarrow v$, et donc $x \rightsquigarrow u \rightarrow \neg v$. Donc, avec $x \rightsquigarrow v$ qui est équivalent à $\neg v \rightarrow \neg x$, on conclut que $x \rightsquigarrow \neg x$. La probabilité que la clause $\neg u \rightarrow \neg v$ apparaisse est au moins $\Theta((n^{2/3}/t^2)^2 tn^{-4/3}) = \Theta(1/t^3)$. Si en plus il apparaît une clause qui contient deux littéraux de la portion d'entrée, alors la formule est insatisfaisable. Ces évènements sont indépendants, donc la probabilité que les deux se produisent est au moins $\Theta(1/t^6)$. Mais avec grande probabilité il y a $\Theta(t^3)$ hourglass, donc avec probabilité au moins $\Theta(1/t^3)$ la formule devient insatisfaisable. En posant $t = 2|\lambda_n|$ on obtient la borne désirée. □

Preuve de la borne supérieure du théorème C.4 en régime sur-critique. On démarre le processus avec $p = (1 - tn^{-1/3})/2n$ (avec $\lambda_0 \leq t \leq \varepsilon_0 n^{1/3}$) où la probabilité qu'il n'y ait pas de hourglass géant est au plus $\exp(-\Theta(t^3))$, et on augmente ensuite jusqu'à $(1 + 2tn^{-1/3})/2n$. Une clause $\neg u \vee \neg v$ telle que u et v sont dans la portion de sortie au début du processus apparaîtra à la fin du processus avec la probabilité au moins $tn^{-4/3}$. Ceci fait que la portion d'entrée du hourglass géant devient partie de l'épine de la formule. La probabilité que ce ne soit pas le cas peut être bornée par $(1 - tn^{-4/3})^{\Theta(tn^{2/3})^2} = \exp(-\Theta(t^3))$. De plus, pour toute clause $u \wedge v$ où u et v sont dans la portion d'entrée du hourglass avant le processus, apparaîtra à la fin avec probabilité au moins $tn^{-4/3}$. Ainsi, la probabilité pour que la formule ne devienne pas insatisfaisable est exponentiellement petite en t^3. On achève la preuve en posant $t = \lambda_n/2$. □