INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **36** (2003) 5719–5730 PII: S0305-4470(03)58571-7

# Information loss in coarse graining of polymer configurations via contact matrices

**Patrik L Ferrari**[1] **and Joel L Lebowitz**[2]

[1] Zentrum Mathematik, Technische Universität München, D-85747 Garching, Germany
[2] Departments of Mathematics and Physics, Rutgers University, Piscataway, NJ, USA

E-mail: ferrari@ma.tum.de and lebowitz@math.rutgers.edu

Received 20 January 2003, in final form 14 April 2003
Published 13 May 2003
Online at stacks.iop.org/JPhysA/36/5719

**Abstract**

Contact matrices provide a coarse grained description of the configuration $\omega$ of a linear chain (polymer or random walk) on $\mathbb{Z}^n$: $\mathcal{C}_{ij}(\omega) = 1$ when the distance between the positions of the $i$th and $j$th steps are less than or equal to some distance $a$ and $\mathcal{C}_{ij}(\omega) = 0$ otherwise. We consider models in which polymers of length $N$ have weights corresponding to simple and self-avoiding random walks, SRW and SAW, with $a$ the minimal permissible distance. We prove that to leading order in $N$, the number of matrices equals the number of walks for SRW, but not for SAW. The coarse grained Shannon entropies for SRW agree with the fine grained ones for $n \leqslant 2$, but differs for $n \geqslant 3$.

PACS numbers: 87.10.+e, 05.50.+q

## 1. Introduction

The use of coarse grained descriptions is essential for systems with many degrees of freedom. The choice of the coarse grained variables is dictated by the nature of the system and by the questions of interest. One is then interested in the amount of information lost in the coarse graining, at least in some statistical sense [2].

In this paper, we shall study this question for simple models of polymers, large molecules consisting of a linear sequence of $N$ monomer units. A reduced description of this system can be based [7, 11, 14, 15] on associating with each polymer configuration a connectivity or contact matrix $\mathcal{C}$, such that $\mathcal{C}_{ij} = 1$ or 0 depending on whether the distance between the positions of the $i$th and $j$th monomers, $\omega(i)$ and $\omega(j)$, is smaller or bigger than a certain specified value $a$,

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{if} \quad |\omega(i) - \omega(j)| \leqslant a \quad i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

This coarse grained (see figure 1) representation of the structure of proteins is often used for numerical studies of protein folding. The very minimalist nature of this representation

  

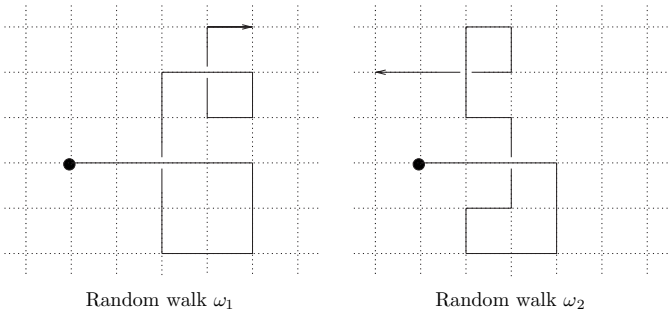


**Figure 1.** The two random walks $\omega_1$ and $\omega_2$ have the same contact matrix because they have the same self-intersections: 2–10 and 13–17.

permits a rapid first search for a protein's native structure in terms of its contact matrix. In fact, knowledge of the contact matrix can predict many features of the vibrational spectra of certain proteins [1, 5]. This makes it important to have information about the relation between the spaces of contact matrices and that of the proteins they represent.

To answer the question of how much information about a polymer is retained by its contact matrix, we consider an idealized version of the geometrical structure of a polymer in which the monomers occupy sites on the $n$-dimensional lattice $\mathbb{Z}^n$, and consecutive monomers are on nearest neighbour lattice sites.

We compare the Shannon entropy after coarse graining, $S_C(N)$, with the Shannon entropy without coarse graining, $S(N)$. To quantify the loss of information in the coarse graining we consider $\delta_N = S_C/S$. We prove, for SRW on $\mathbb{Z}^2$, that $\delta_N$ goes to one as $N$ becomes large, showing that the relative loss of information $(S - S_C)/S$ vanishes. This is a consequence of the recurrence of SRW on $\mathbb{Z}^2$. Moreover, we provide some bounds on finite size corrections (see (3.1*b*)). On the other hand, for SRW in $\mathbb{Z}^n$, $n \geqslant 3$, and for SAW in $\mathbb{Z}^n$, $n \geqslant 2$, $\delta_N$ remains strictly less than one, which means that the loss of information due to the coarse graining becomes substantial (see (3.2*b*) and theorem 3).

We also consider the problem addressed already in [14], i.e., how the number of different *physical* contact matrices $W(N)$ depends on the polymer length $N$. It was shown there analytically that $W(N)$ increases exponentially in $N$, and numerically that the growth exponent $\gamma_N$ is strictly less than the growth exponent for the total number of SAW. In our work, we give a rigorous proof of this (see theorem 3). We also consider the same problem for SRW. Surprisingly, the growth exponent for the contact matrices is now the same as for the number of SRW in all dimensions. The reason for this is that the probability distribution of the number of distinct visited sites (divided by $N$) has a left tail which does not decay exponentially fast in $N$. To conclude, we provide lower bounds on $\gamma_N$, relevant to the finite size behaviour (see (3.1*a*) and (3.2*a*)).

The outline of the rest of the paper is as follows. In section 2, we introduce the model, the relevant quantities and the studied examples. The main results are presented and briefly discussed in section 3. Sections 4–6 are devoted to the proof of the main results.

## 2. Preliminaries

We define more precisely the quantities and the examples which will be studied. $\Omega_N$ is the set of all polymers containing $N + 1$ monomers, where the configuration of such a polymer is specified by $\omega_N = (\omega(0), \omega(1), \ldots, \omega(N))$ with $\omega(0) \equiv 0$ and $\omega(i+1) - \omega(i) = \pm e_\alpha$, where

$e_\alpha$ is one of the unit directions on $\mathbb{Z}^n, \alpha = 1, \ldots, n$. Let there be given some probability distribution $\mathbb{P}(\omega)$ on $\Omega$. (We shall drop the subscript $N$ whenever possible.) The contact matrices $\mathbf{C} = \{\mathcal{C}(\omega)\}_{\omega\in\Omega}$ partition $\Omega$ into sets $\Omega_C = \{\omega : \mathcal{C}(\omega) = C\}$, with

$$\deg \mathcal{C} = |\Omega_C| \tag{2.1}$$

the number of configurations $\omega \in \Omega_C$. The probability of $\omega$ being in $\Omega_C$ is then

$$\mathbb{P}(C) = \sum_{\omega\in\Omega_C} \mathbb{P}(\omega). \tag{2.2}$$

To measure the information lost in the coarse graining, we may compare the Shannon entropy $S_C$ of the coarse grained measure $\mathbb{P}(C)$ with the fine grained entropy $S$,

$$S = -\sum_{\omega\in\Omega} \mathbb{P}(\omega) \ln \mathbb{P}(\omega). \tag{2.3}$$

We then have

$$S_C = -\sum_{C\in\mathbf{C}} \mathbb{P}(C) \ln \mathbb{P}(C) = -\sum_{C\in\mathbf{C}} \sum_{\omega\in\Omega_C} \mathbb{P}(\omega) \ln\left(\frac{\mathbb{P}(C)}{\mathbb{P}(\omega)}\mathbb{P}(\omega)\right) = S - \hat{S}_C \tag{2.4}$$

where

$$\hat{S}_C = -\sum_{C\in\mathbf{C}} \mathbb{P}(C) \sum_{\omega\in\Omega_C} \mathbb{P}(\omega|C) \ln \mathbb{P}(\omega|C) \tag{2.5}$$

with

$$\mathbb{P}(\omega|C) = \mathbb{P}(\omega)/\mathbb{P}(C) \qquad \text{for} \quad \omega \in \Omega_C \tag{2.6}$$

is the conditional probability of $\omega$ given that it is in $\Omega_C$. We can thus think of $\hat{S}_C$ as the average 'conditional entropy' relative to $\mathcal{C}$. Since $\hat{S}_C \geqslant 0$, we clearly have $S_C \leqslant S$, and $S - S_C$ is then a measure of information lost in the coarse graining [2]. The question is how much. In particular, we may ask how does

$$\delta_N = S_C/S \tag{2.7}$$

behave as $N \to \infty$.

Before answering this question, we note that

$$S_C \leqslant \bar{S}_C = \ln|\mathbf{C}| \tag{2.8}$$

where $\bar{S}_C$ is the entropy of the distribution $\bar{\mathbb{P}}(C)$ which assigns equal weight to all $C \in \mathbf{C}$, i.e., $\bar{\mathbb{P}}(C) = W(N)^{-1}$, with $W(N) \equiv |\mathbf{C}|$ the total number of different coarse grained components, i.e., contact matrices. We may then also define

$$\gamma_N = \bar{S}_C/S \tag{2.9}$$

and by (2.8)

$$\delta_N \leqslant \gamma_N \leqslant 1. \tag{2.10}$$

The last inequality is obtained by replacing $\mathbb{P}(C)$ by $\bar{\mathbb{P}}(C)$ in (2.4)–(2.6) and using that the corresponding $\hat{S}_C$ is positive.

So far everything is completely general. We shall now specialize to the case where all permissible configurations $\omega$, i.e., all those for which $\mathbb{P}(\omega) \neq 0$, have the same probability. Then

$$W(N) = |\mathbf{C}| = \sum_{\omega\in\Omega} 1/\deg \mathcal{C}(\omega) = |\Omega|\mathbb{E}(\deg \mathcal{C})^{-1} \tag{2.11}$$

where $\mathbb{E}$ is the expectation value with respect to the relevant uniform distribution.

The examples we shall consider here are

(1) The weights are those of simple symmetric random walks (SRW) on $\mathbb{Z}^n$, i.e., $|\Omega| = (2n)^N$ and $\mathbb{P}(\omega) = (2n)^{-N}$ for all $\omega$.
(2) The polymers behave like self-avoiding walks (SAW) on $\mathbb{Z}^n$, i.e., the configuration space $\Omega$ consists of all $\omega$ s.t. $\omega(i) \neq \omega(j)$ for $i \neq j$, and $\mathbb{P}(\omega) = |\Omega|^{-1}$, where $|\Omega| \sim \mu_{\mathrm{SAW}}^N$ is the number of SAW on $\mathbb{Z}^n$ of length $N$. SAW model the steric exclusion effects of the monomers and are frequently used as a model for polymers [4, 9].
(3) The chains behave like bond self-avoiding walks (BAW) on $\mathbb{Z}^n$ in which case $\Omega$ consists of all $\omega$ such that the pair $[\omega(i), \omega(i+1)] \neq [\omega(j), \omega(j \pm 1)]$ for $i \neq j$, and $\mathbb{P}(\omega) = |\Omega|^{-1}$, $|\Omega| \sim \mu_{\mathrm{BAW}}^N$, the number of BAW [12].

Note that for uniform distributions, $S$ is just the logarithm of the total number of configurations, i.e., $S = \ln|\Omega|$, so $\gamma_N$ is just the ratio of the logarithms of the numbers of contact matrices and random walks.

The behaviour of $\gamma_N$ was studied in [14] for the case of SAW, with

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{if} \quad |\omega(i) - \omega(j)| = 1 \quad |i-j| > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

(The inequality can in fact be made an equality here since 1 is the minimal distance between $\omega(i)$ and $\omega(j)$ for $i \neq j$.) Numerical studies [14] for $n = 2$ indicated that $\gamma_N$ remains strictly less than 1 in the limit $N \to \infty$. It is then natural to ask whether the same is true for the SRW when we again define $\mathcal{C}_{ij}(\omega) = 1$ when $\omega(i)$ and $\omega(j)$ are as close as they can be, i.e., when $a$ in (1.1) is set equal to zero

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{for} \quad |\omega(i) - \omega(j)| = 0 \quad i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

Note that for this case, $W(N)$ satisfies

$$W(N_1 + N_2) \geqslant W(N_1)W(N_2). \tag{2.14}$$

Since for SRW $\gamma_N = \frac{\ln W(N)}{N \ln(2n)}$, it follows from (2.14) that $\gamma_N$ is monotone non-decreasing in $N$ and thus that $\lim_{N\to\infty} \gamma_N$ exists: remember $\gamma_N \leqslant 1$.

In the present work, we prove some results about $\delta_N$ and $\gamma_N$ for all the above examples. (Some of these generalize readily to other uniform distributions.)

## 3. Main Results

**Theorem 1.** *For SRW on $\mathbb{Z}^2$, there exist constants $\kappa, \kappa_1, \kappa_2 > 0$ such that for large N,*

$$\gamma_N \geqslant 1 - \frac{\kappa \ln N}{N^{1/3}} \tag{3.1a}$$

$$1 - \frac{\kappa_1}{\ln N} \leqslant \delta_N \leqslant 1 - \frac{\kappa_2}{(\ln N)^2}. \tag{3.1b}$$

*Consequently, $\gamma_N$ and $\delta_N \to 1$, as $N \to \infty$.*

**Theorem 2.** *For SRW on $\mathbb{Z}^n$, $n \geqslant 3$, there exist constants $\kappa_n, \kappa_n' > 0$ such that for N large enough,*

$$\gamma_N \geqslant 1 - \frac{\kappa_n}{N^{2/(n+2)}} \tag{3.2a}$$

$$\delta_N \leqslant 1 - \kappa_n'. \tag{3.2b}$$

*Hence, $\gamma_N \to 1$, as $N \to \infty$ while $\limsup_{N\to\infty} \delta_N < 1$.*

**Theorem 3.** *For SAW and BAW on* $\mathbb{Z}^n, n \geqslant 2, \limsup_{N\to\infty} \gamma_N < 1$ *and ipso-facto* $\limsup_{N\to\infty} \delta_N < 1$.

We are indebted to Harry Kesten for the key idea in the proof of theorem 3 for SAW. The extension to BAW is straightforward.

Intuitively, we expect that the larger the number of intersections, the more information is contained in the contact matrices. It is therefore not surprising that for recurrent RW, such as SRW in $n = 2$, both $\gamma_N$ and $\delta_N$ would go to 1. (For the degenerate case $n = 1$, there are, for all the cases, just twice as many random walks as contact matrices corresponding to whether the first step is to the right or the left.) One expects however that for RW which have a strong tendency to spread out such as SAW in $n \geqslant 2$, and SRW in $n \geqslant 3$, the contact matrices lose too much information. This is indeed reflected in $\delta_N < 1$ for SAW in $n \geqslant 2$, and SRW in $n \geqslant 3$. The same is true for $\gamma_N$ for SAW. Surprisingly, however $\gamma_N \to 1$ for SRW in all dimensions. The reason for this, as we shall see, is that the probability that a SRW of length $N$ in $\mathbb{Z}^n$ visits $R_N \leqslant \epsilon N$ distinct sites goes to zero slower than exponentially when $N \to \infty$ for any fixed $\epsilon > 0$.

The outline of the rest of the paper is as follows. In section 4, we first present some general inequalities and then prove the results about $\gamma_N$ for SRW. In section 5, we give some bounds on the degeneracy of SRW and then prove the results about $\delta_N$ for SRW. In section 6, we prove theorem 3 for SAW and BAW.

## 4. Proof of results for $\gamma_N$

*Some inequalities.* Using Jensen's inequality on the average over $\Omega_C$, gives

$$\sum_{\omega\in\Omega_C} \frac{1}{\deg \mathcal{C}} (\mathbb{P}(\omega) \deg \mathcal{C}) \ln(\mathbb{P}(\omega) \deg \mathcal{C}) \geqslant \sum_{\omega\in\Omega_C} \mathbb{P}(\omega) \ln\left( \sum_{\omega\in\Omega_C} \mathbb{P}(\omega) \right)$$
$$= \mathbb{P}(C) \ln \mathbb{P}(C). \tag{4.1}$$

Writing now

$$S = -\sum_{C\in\mathbf{C}} \sum_{\omega\in\Omega_C} \mathbb{P}(\omega)(\ln(\mathbb{P}(\omega) \deg \mathcal{C}) - \ln \deg \mathcal{C}) \tag{4.2}$$

we obtain[1]

$$S \leqslant S_C + \mathbb{E}(\ln \deg \mathcal{C}) \tag{4.3}$$

which yields

$$1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{S} \leqslant \delta_N \leqslant 1. \tag{4.4}$$

We next give an upper bound for the degeneracy of the contact matrices defined in (2.13). This is the key to our results for SRW. Note that for the examples considered here, the first inequality in (4.4) is indeed an equality.

Define the *range* $R_N$ of a SRW with length $N$ to be *the number of distinct sites visited by the walk*.

**Lemma 4.** *Let* $\omega$ *have a range* $R_N = M$ *and let* $\mathcal{C}(\omega)$ *be its contact matrix. Then*

$$\deg \mathcal{C}(\omega) \leqslant (2n)^M. \tag{4.5}$$

[1] The right side of (4.3) is the maximum of the entropy over all measures, $\mu(\omega)$ such that $\mu(\mathcal{C}) = \mathbb{P}(\mathcal{C})$.

**Proof.** The contact matrix $\mathcal{C}(\omega)$ has $N+1-M$ columns with '1's in the upper triangular part, because we have $N+1-M$ intersections. Let us now construct all random walks $\omega'$ such that $\mathcal{C}(\omega') = \mathcal{C}(\omega)$. Consider now $\omega'(k)$ with $k \neq 0$, then there are two possible cases:

1. there exists an $i < k$ such that $\mathcal{C}_{ik}(\omega) = 1$;
2. for all $i < k$, $\mathcal{C}_{ik}(\omega) = 0$.

In the first case, $\omega'(k) = \omega'(i)$ and therefore we have only one choice for it. In the second case, the $k$th step will occupy a place which was never occupied before. Therefore, we have *at most* $2n$ possible choices.

For a $\omega'$ with a contact matrix $\mathcal{C}(\omega') = \mathcal{C}(\omega)$, there are $M-1$ steps for which we are in the second case because the starting point is fixed at the origin and $N+1-M$ steps for which we are in the first one. Therefore, there are *at most* $(2n)^{M-1} \leqslant (2n)^M$ different $\omega'$ satisfying $\mathcal{C}(\omega') = \mathcal{C}(\omega)$, i.e., $\deg \mathcal{C}(\omega) \leqslant (2n)^M$. □

**Proof of (3.1*a*).** We first note that $\delta_N, \gamma_N \to 1$ for general recurrent RW, which includes SRW on $\mathbb{Z}^2$. For such walks $\mathbb{E}(R_N)/N \to 0$ as $N \to \infty$, therefore using (4.4) and lemma 4,

$$\delta_N \geqslant 1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{N \ln 2n} \geqslant 1 - \frac{\mathbb{E}(R_N)}{N} \to 1 \quad \text{as} \quad N \to \infty. \tag{4.6}$$

To prove (3.1*a*) consider the subset $\Omega^\alpha$ of SRW on $\mathbb{Z}^2$ defined as

$$\Omega^\alpha = \{\omega \in \Omega_N \text{ s.t. } \omega(k \times 4[N^\alpha]) = 0, k = 0, 1, \ldots, k_{\max}\} \tag{4.7}$$

where $k_{\max}$ is the largest integer $k$ such that $k4[N^\alpha] \leqslant N$. Let us take $0 < \alpha < 1/2$. Each $\omega \in \Omega^\alpha$ returns to the origin after $4[N^\alpha]$ steps, it is therefore contained in a cube of edge length $4[N^\alpha]$ (except eventually for the last $2[N^\alpha]$ steps). Using Stirling formula we have, for some $K > 0$,

$$\mathbb{P}(\omega(4M) = 0) \geqslant \frac{(4M)!}{(M!)^4 4^{4M}} \geqslant K M^{-3/2} \tag{4.8}$$

because $\{\omega(4M) = 0\} \supset \{\omega(4M) = 0 \text{ with } M \text{steps in each direction}\}$. Then

$$\mathbb{P}(R_N \leqslant N^\beta \equiv 4^2[N^\alpha]^2 + 2[N^\alpha]) \geqslant \mathbb{P}(\omega \in \Omega^\alpha) \geqslant (K/[N^\alpha]^{3/2})^{N/4[N^\alpha]}. \tag{4.9}$$

Therefore combining (2.11), (4.5) and (4.9) for $n = 2$, we obtain

$$W(N) \geqslant 4^N \mathbb{P}(R_N \leqslant N^\beta)/4^{(N^\beta)} \tag{4.10}$$

which implies

$$1 \geqslant \gamma_N \geqslant 1 - \kappa \frac{\ln N}{[N^\alpha]} - \frac{4^2[N^\alpha]^2 + 2[N^\alpha]}{N} \tag{4.11}$$

for a $\kappa > 0$. For $\alpha = 1/3$, the rhs of (4.11) is optimized and the term with the logarithm dominates the last one. This proves the bound for $n = 2$. □

**Proof of (3.2*a*).** For $n > 2$, $\mathbb{P}(R_N = X) \sim \exp(-aN/X^{2/n})$ when $X \to \infty$, $\frac{X}{N} \to 0$ (see [13] and pp 88–92 of [18]). Therefore for $\alpha \in (0, 1)$, $\mathbb{P}(R_N = N^\alpha) \sim \exp(-aN^{1-2\alpha/n})$. But for an $\omega$ with $R_N = M$, $\deg \mathcal{C}(\omega) \leqslant (2n)^M$, (see lemma 4). Therefore, restricting the sum in (2.11) to $\omega \in \Omega_N^\alpha$, we have $W(N) \geqslant (2n)^N \mathbb{P}(R_N = N^\alpha)/(2n)^{N^\alpha}$, since the numerator is just the number of terms in that sum. This implies, for large $N$,

$$1 \geqslant \gamma_N \geqslant 1 - \sigma(N, \alpha) \tag{4.12}$$

where $\sigma(N, \alpha) = N^{\alpha-1} + aN^{-2\alpha/n}/\ln 2n$. Choosing $\alpha \in (0, 1)$ which minimizes $\sigma(N, \alpha)$ for large $N$, we obtain $\alpha - 1 = -2/(n+2)$. Taking $\kappa_n = (\ln 2n + a)/\ln 2n$ completes the proof. □

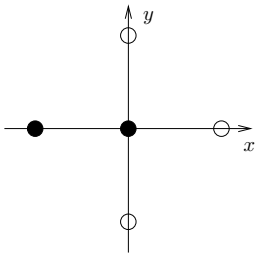


**Figure 2.** The pattern $Q$. The visited sites of $Q$ are black.

This theorem implies that for $N$ large, we have (up to smaller corrections),

$$W(N) \geqslant (2n)^N (2n)^{-\kappa_n N^{n/(n+2)}}. \tag{4.13}$$

There exists also an upper bound on $\gamma_N$ which depends on the decrease of $\mathbb{P}(R_N/N < \varepsilon)$.

**Proposition 5.** *For all fixed $\varepsilon > 0$, there exists a constant $\kappa' > 0$ such that for N large enough,*

$$\gamma_N \leqslant 1 - \kappa' \frac{|\ln \mathbb{P}(R_N/N \leqslant \varepsilon)|}{N}. \tag{4.14}$$

The outline of the proof will be given in the appendix.

## 5. Bounds on the degeneracy of SRW

### *5.1. SRW on $\mathbb{Z}^2$*

For SRW on $\mathbb{Z}^2$, $R_N \sim \frac{\pi N}{\ln N}$, more precisely (see, e.g. [16]),

$$\mathbb{E}(R_N) = \frac{\pi N}{\ln 8N} \left(1 + \mathcal{O}\left(1/\ln N\right)\right). \tag{5.1}$$

Next we apply a result of van Wijland *et al* [16]. Let the *support of $\omega$* be defined to be *the set of points visited by $\omega$*. Consider two finite disjoint sets of lattice points $A_u$ and $A_v$. The 'pattern' centred at $\mathbf{x}$ associated with the sets $A_u$ and $A_v$ is a configuration of $|A_v|$ visited sites $\mathbf{x}+\mathbf{z}$, $\mathbf{z} \in A_v$ and of $|A_u|$ unvisited sites $\mathbf{x}+\mathbf{z}'$, $\mathbf{z}' \in A_u$. We say that the pattern appears in the support of $\omega$ at $\mathbf{x}$ if the lattice points $\mathbf{x}+\mathbf{z}$, $\mathbf{z} \in A_v$, are in the support of $\omega$ and the lattice points $\mathbf{x}+\mathbf{z}'$, $\mathbf{z}' \in A_u$, are not in the support of $\omega$. The numbers of times that a pattern appears in the support of $\omega$ is then the number of different $\mathbf{x} \in \mathbb{Z}^n$ such that it appears at $\mathbf{x}$.

Let us consider the 'pattern $Q$' defined as the set composed of the following sets $A_v$ and $A_u$: $A_v(Q) = \{(0,0), (-1,0)\}$ and $A_u(Q) = \{(1,0), (0,-1), (0,1), (-1,1)\}$ (see figure 2).

Let $Q_N = Q_N(\omega)$ be the number of times that $Q$ appears in the support of $\omega$. Then using [16], $\mathbb{E}(Q_N) = \frac{\pi^2 N}{(\ln 8N)^2} m_1 + \mathcal{O}\big(\frac{N}{(\ln 8N)^3}\big)$ where $m_1 = m_1(Q)$ is a constant, and $Q_N - \mathbb{E}(Q_N) \simeq \frac{2\mathcal{A}}{\ln 8N} \mathbb{E}(Q_N)\gamma(N)$ where $\gamma(N)$ is a random variable (Varadhan's renormalized local time of self-intersections, see [10]) with mean 0 and variance 1. $\mathcal{A}$ is a constant given in [16] whose value is $\sim$1.3034. We computed $m_1$ finding $m_1 \cong 2.78 \times 10^{-3}$.

These results imply the following proposition.

**Proposition 6.** *For simple random walks on $\mathbb{Z}^2$, there exists a $\nu > 0$ such that*

$$\lim_{N\to\infty} \mathbb{P}\big(\deg \mathcal{C}(\omega) \geqslant e^{\nu N/(\ln N)^2}\big) = 1. \tag{5.2}$$

**Proof.** Suppose that the pattern $Q$, centred at $\zeta \in \mathbb{Z}^2$, exists in the support of a random walk $\omega$. Let us consider the following transformation:

$$\begin{aligned} T_\zeta : \Omega_N &\longmapsto \Omega_N \\ \omega &\longrightarrow T_\zeta(\omega) = \begin{cases} \omega(i) & \text{if} \quad \omega(i) \neq \zeta \\ \zeta + (-1, 1) & \text{if} \quad \omega(i) = \zeta. \end{cases} \end{aligned} \tag{5.3}$$

In other words, we exchange the points $\zeta$ and $\zeta + (-1, 1)$. This application does not change the contact matrix of the random walk, because $\zeta + (-1, 1)$ is connected only with $\zeta + (-1, 0)$. We have to prove that the probability of having the pattern $Q$ in the support of a random walk at least $M = \nu N/(\ln N)^2$ times goes to 1 as $N \to \infty$. A RW with $M$ times the pattern $Q$ appearing in its support is at least $2^M$ times degenerate: we can apply or not apply $T_\zeta$ independently for each $\zeta$ such that $Q$ appears in the support of $\omega$ (centred in $\zeta$).

We want an upper bound of $\mathbb{P}\big(Q_N < \alpha \frac{\mu N}{(\ln N)^2}\big)$ for $\alpha \in (0, 1)$ and $\mu = m_1 \pi^2$. For each $k > 0$ and $N$ large enough,

$$\mathbb{P}\left( Q_N < \alpha \frac{\mu N}{(\ln N)^2} \right) \leqslant \mathbb{P}\left( Q_N - \mathbb{E}(Q_N) \leqslant -k a_Q \frac{N}{(\ln 8N)^3} \right) \tag{5.4}$$

with $a_Q = 2\mu\mathcal{A}$. In fact, for $N$ large enough, $\mathbb{E}(Q_N) = \frac{\mu N}{(\ln 8N)^2} + \mathcal{O}(N/(\ln 8N)^3)$ and therefore for each $\alpha < 1$, $\frac{\mu N}{(\ln 8N)^2} + \mathcal{O}(N/(\ln 8N)^3) - k a_Q \frac{N}{(\ln 8N)^3} \geqslant \alpha \frac{\mu N}{(\ln N)^2}$. Thus

$$\begin{aligned} \mathbb{P}\left( Q_N < \alpha \frac{\mu N}{(\ln N)^2} \right) &\leqslant \mathbb{P}\left( Q_N - \mathbb{E}(Q_N) \leqslant -k a_Q \frac{N}{(\ln 8N)^3} \right) \\ &\leqslant \frac{\mathbb{E}\left(Q_N - \mathbb{E}(Q_N)\right)^2}{k^2 a_Q^2 \frac{N^2}{(\ln 8N)^6}} \overset{N\to\infty}{\longrightarrow} \frac{1}{k^2}. \end{aligned} \tag{5.5}$$

Therefore for each $\alpha \in (0, 1)$, we have $\forall k > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left( Q_N < \alpha \frac{\mu N}{(\ln N)^2} \right) \leqslant \frac{1}{k^2}. \tag{5.6}$$

This implies that for all $\alpha \in (0, 1)$, $\lim_{N\to\infty} \mathbb{P}\big(Q_N \geqslant \alpha \frac{\mu N}{(\ln N)^2}\big) = 1$. All the random walks with a number of pattern $Q$ in their support more than $\alpha \frac{\mu N}{(\ln N)^2}$ are more degenerate than $\mathrm{e}^{\alpha\mu N \ln 2/(\ln N)^2}$. Then, since $\alpha \in (0, 1)$, for all choice of $\nu < \mu \ln 2$, we have (5.2). □

### *5.2. SRW on $\mathbb{Z}^n$, $n \geqslant 3$*

Let us define, for $n = 3$, the pattern $P$ as consisting of a set $A_v$ of visited sites and a set $A_u$ of unvisited sites as follows: $A_v(P) = \{(0, 0, 0), (-1, 0, 0)\}$ and $A_u(P) = \{(1, 0, 0), (0, 1, 0), (0, -1, 0), (0, 0, 1), (0, 0, -1), (-1, 1, 0)\}$.

Let $P_N = P_N(\omega)$ be the number of times that $P$ appears in the support of $\omega$. Then using [17] we have $\mathbb{E}(P_N) = m_1 N + \mathcal{O}(\sqrt{N})$ and $P_N - \mathbb{E}(P_N) \simeq a_P \sqrt{N \ln N} \eta(N)$ where $m_1 = 2.5 \times 10^{-3}$, $a_P = 1.2 \times 10^{-2}$ and $\eta(N)$ is a random variable with normal distribution $\mathcal{N}(0, 1)$.

**Proposition 7.** *For simple random walks on $\mathbb{Z}^3$, there exists a $\nu > 0$ such that*

$$\lim_{N\to\infty} \mathbb{P}(\deg \mathcal{C}(\omega) \geqslant \mathrm{e}^{\nu N}) = 1. \tag{5.7}$$

**Proof.** The proof is very close to the one of proposition 6. This time we exchange the sites $\zeta$ and $\zeta + (-1, 1, 0)$ (if $P$ appears centred in $\zeta$) and we prove that $\forall k > 0$ and $N$ large enough $\mathbb{P}(P_N < \alpha m_1 N) \leqslant \frac{1}{k^2}$ if $\alpha < 1$. Then for all choice of $\nu < m_1 \ln 2$ (5.7) holds. □

In dimension $n \geqslant 4$, the same result (with a different value of $m_1$) is expected to hold. In fact a similar pattern in $n \geqslant 4$ has $\mathbb{E}(P_N) = m_1 N + \mathcal{O}(\ln N)$ in $n = 4$, $\mathbb{E}(P_N) = m_1 N + \mathcal{O}(1)$ in $n \geqslant 5$ and $\mathbb{E}(P_N - \mathbb{E}(P_N))^2 = K_P N + o(N)$, (see [17]). The only point that one should prove for $n \geqslant 4$ is that $m_1 \neq 0$.

### *5.3. Proof of (3.1b) and (3.2b)*

The previous results on the degeneracy lead to the following results.

**Proof of (3.1*b*).** Using propositions 6 and (5.1), we obtain bounds on $\delta_N$, for $n = 2$.

$$\delta_N \leqslant 1 - \mathbb{P}\big(\deg \mathcal{C} \geqslant \mathrm{e}^{\nu N/(\ln N)^2}\big)\frac{\nu}{\ln 4 (\ln N)^2}. \tag{5.8}$$

By proposition 6, there exists a $\nu > 0$ such that $\mathbb{P}\big(\deg \mathcal{C} \geqslant e^{\nu N/(\ln N)^2}\big) \to 1$ as $N \to \infty$. This gives

$$\lim_{N\to\infty} (1 - \delta_N)(\ln N)^2 \geqslant \nu/\ln 4. \tag{5.9}$$

Consequently for $\kappa_2 < \nu/\ln 4$ and $N$ large enough, $\delta_N \leqslant 1 - \frac{\kappa_2}{(\ln N)^2}$.

On the other hand,

$$\delta_N \geqslant 1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{N \ln 2n} \geqslant 1 - \frac{\mathbb{E}(R_N)}{N}. \tag{5.10}$$

Then

$$\lim_{N\to\infty} (1 - \delta_N) \ln N \leqslant \lim_{N\to\infty} \mathbb{E}(R_N)\frac{\ln N}{N} = \pi \tag{5.11}$$

consequently for $\kappa_1 > \pi$ and $N$ large enough, $\delta_N \geqslant 1 - \kappa_1/\ln N$. □

**Proof of (3.2*b*).** For all $\nu > 0$,

$$\delta_N \leqslant 1 - \frac{\mathbb{P}(\deg \mathcal{C} \geqslant \mathrm{e}^{\nu N})}{\ln(2n)}\nu. \tag{5.12}$$

By proposition 7, there exists, for $n \geqslant 3$, a $\nu > 0$ such that $\mathbb{P}(\deg \mathcal{C} \geqslant \mathrm{e}^{\nu N}) \to 1$ as $N \to \infty$. Therefore for $N$ large enough

$$\delta_N \leqslant 1 - \frac{\nu/2}{\ln(2n)} < 1 \tag{5.13}$$

and $\limsup_{N\to\infty} \delta_N < 1$. □

## 6. Proof of theorem 3

The contact matrix of $\omega$ is now defined by (2.12).

Let us consider the case of SAW. We introduce some notation: we consider a cube $D = \{x \in \mathbb{Z}^n \text{ s.t. } c_i \leqslant x^{(i)} \leqslant c_i + b, 1 \leqslant i \leqslant n\}$ for some $c = (c_1, \ldots, c_n) \in \mathbb{Z}^n$ with its boundary $\partial D = \{x \in \mathbb{Z}^n \text{ s.t. } x^{(i)} = c_i + b \text{ or } x^{(i)} = c_i, 1 \leqslant i \leqslant n\}$. A path $P$ is a SAW of finite length, say $k$, starting at the origin, i.e., $P = \{X_i(P), 0 \leqslant i \leqslant k\}$ with $X_0(P) = 0$.

We consider only paths such that there exists a cube $D$ with $X_0(P) = 0$ and $X_k(P)$ two of its vertices and $X_j(P) \in D$ for all $0 \leqslant j \leqslant k$. We say that $(P, D)$ occurs at the $r$th step $\omega$ if

(1) $X_{r+j}(\omega) - X_r(\omega) = X_j(P)$ for all $j = 0, \ldots, k$ and
(2) $\omega$ does not occupy any other points of $D$.

$\chi_N(j, (P, D))$ is the number of $\omega \in \Omega_N$ such that $(P, D)$ occurs *at most* at $j$ steps.

**Theorem 8 (Kesten pattern theorem [8]).** *Let $P$ be a SAW and $D$ a cube such that $D$ has 0 and $X_k(P)$ as two of its vertices and contains $P$. Then*

$$\limsup_{N\to\infty}\left(\frac{\chi_N(aN,(P,D))}{|\Omega_N|}\right)^{1/N} < 1 \qquad \text{for some } a > 0 \tag{6.1}$$

*where $|\Omega_N|$ is the total number of SAW of length $N$.*

It is known that $|\Omega_N| \simeq \mu_{\mathrm{SAW}}^N$ with $\mu_{\mathrm{SAW}} > 0$.

**Proof of theorem 3.** Let us take $b > 2$ and consider a SAW path of length $k + 2n$ constructed as follows. The firsts $n$ steps of $P$ connect the points $(0,\dots,0)$ and $(1,\dots,1)$. The following $k$ steps connect the points $(1,\dots,1)$ and $(b-1,\dots,b-1)$ with a SAW remaining always in $D\setminus\partial D$. The last $n$ steps of $P$ connect the points $(b-1,\dots,b-1)$ and $(b,\dots,b)$. Let us divide the set $\Omega_N$ into a sum of two disjoint parts: $\Omega_N = \Omega_N^a \cup \left(\Omega_N^a\right)^c$ where $\Omega_N^a = \{\omega \in \Omega_N \text{ s.t. } (P,D) \text{ occurs at most } a\,N \text{ times}\}$ and $\left(\Omega_N^a\right)^c$ its complementary set. It follows from theorem 8 that

$$\exists \zeta > 0 \qquad \text{s.t.} \quad \mathbb{P}\left(\omega \in \Omega_N^a\right) \leqslant \mathrm{e}^{-\zeta N}.$$

Let us take a $\omega \in \left(\Omega_N^a\right)^c$. Then $(P,D)$ occurs at least $aN$ times in $\omega$. Consider an occurrence of $(P,D)$ in the piece of $P$ between its $t$th and its $(t+k+2n)$th steps. We apply an axis rotation of $2\pi/n$ degrees to the cube $D\setminus\partial D$, where the axis is its diagonal of direction $(1,\dots,1)$. This transformation does not change the contact matrix, and we can apply it $n$ times obtaining each time a different SAW. For the chosen $\omega$, it can be done independently in at least $aN$ different places, therefore the corresponding contact matrix is at least $n^{aN}$ times degenerate.

Now we have an upper bound for the total number of contact matrices:

$$\begin{aligned} W(N) &\leqslant \mathbb{P}\left(\omega \in \Omega_N^a\right)|\Omega_N| + \mathbb{P}\left(\omega \in \left(\Omega_N^a\right)^c\right)|\Omega_N| n^{-aN} \\ &\leqslant (\mathrm{e}^{-\zeta N} + \mathrm{e}^{-Na\ln n})|\Omega_N|. \end{aligned} \tag{6.2}$$

Defining $\alpha_M = \max\{\zeta, a\ln n\} > 0$ and $\alpha_m = \min\{\zeta, a\ln n\} > 0$, we obtain

$$\begin{aligned} \limsup_{N\to\infty}\gamma_N &\leqslant \lim_{N\to\infty}\frac{\ln\left(\mathrm{e}^{-\alpha_m N}\left(1+\mathrm{e}^{-\frac{\alpha_M}{\alpha_m}N}\right)\right)+\ln|\Omega_N|}{\ln|\Omega_N|} \\ &= 1 - \frac{\alpha_m}{\ln\mu_{\mathrm{SAW}}} < 1. \end{aligned} \tag{6.3}$$

□

Now we consider the case of BAW. We introduce some notation: we consider a cube $D$ as for SAW and a cube $D^1 = \{x \in \mathbb{Z}^n \text{ s.t. } c_i - 1 \leqslant x^{(i)} \leqslant c_i + b + 1, 1 \leqslant i \leqslant n\}$ for some $c = (c_1,\dots,c_n) \in \mathbb{Z}^n$. In this case, a path $P$ is a BAW instead of a SAW with the same conditions as for SAW. We consider only the paths such that there exists a cube $D$ with $X_0(P) = 0$ and $X_k(P)$ two of its vertices and $X_j(P) \in D^1$ for all $0 \leqslant j \leqslant k$. We say that $(P,D)$ occurs at the $r$th step $\omega$ if

1. $X_{r+j}(\omega) - X_r(\omega) = X_j(P)$ for all $j = 0,\dots,k$ and
2. $\omega$ does not occupy any other points of $D$.

$\chi_N(j,(P,D))$ is the number of $\omega \in \Omega_N$ such that $(P,D)$ occurs *at most* at $j$ steps. Theorem 8 holds also for BAW [3].

**Proposition 4.** *For BAW*

$$\limsup_{N\to\infty} \gamma_N < 1. \tag{6.4}$$

**Proof.** The proof is identical to the one of theorem 3. □

**Remark 10.** As noticed by Kesten in [8], theorem 8 could be proven also for other lattices in almost the same way, therefore theorem 3 should hold for other lattices than $\mathbb{Z}^n$.

**Acknowledgments**

We thank H Kesten for supplying us with the argument needed to prove theorem 3 for SAW, and S Goldstein for useful discussions. Research supported by NSF Grant DMR 98-13268, AFOSR Grant AF 49620-01-1-0154, and DIMACS and its supporting agencies, the NSF under contract no STC-91-19999 and the N.J. Commission on Science and Technology. Work of P L Ferrari partially supported by the Swiss fellowship Sunburst-Fonds.

## Appendix. Outline of the proof of proposition 5

Let $I_N = N+1-R_N$ be the number of intersections. Consider an interval $J = [k_0, k_1] \subset [0, 1]$ and the subset $\Lambda_N(J) = \{\omega \in \Omega_N \text{ s.t. } I_N(\omega)/N \in J\}$. We define the mean degeneracy on $\Lambda_N(J)$ by $\langle \deg \mathcal{C} \rangle_J = |\Lambda_N(J)|/W(N)_J$, where $W(N)_J$ is the number of contact matrices corresponding to RW with $I_N/N \in J$. We set $d(J) = \liminf_{N\to\infty} \frac{1}{N} \ln \langle \deg \mathcal{C} \rangle_J$.

**Theorem 11.** *For SRW on* $\mathbb{Z}^n$, $n \geqslant 2$, *and* $\pi' = \lim_{N\to\infty} \mathbb{E}(I_N)/N$,

$$d(J = [k_0, k_1]) > 0 \qquad \textit{for all} \quad k_0 < \pi' \quad \textit{and} \quad k_0 < k_1 < 1. \tag{A.1}$$

For a $\omega \in \Omega_N$, let us define $F(\omega)$ to be the number of loops of length 4 which do not intersect the remaining part of $\omega$ (called 'free-4-loops').

**Proposition 12.** *Let J be as in theorem 11. Then there exists an* $\alpha_J > 0$ *such that*

$$\beta_J = \liminf_{N\to\infty} -\frac{1}{N} \ln \mathbb{P}\{\omega \in \Lambda_N(J) \text{ s.t. } F(\omega) \leqslant \alpha_J N\} > 0. \tag{A.2}$$

**Proof of theorem 11.** For $k_0 < \pi'$ and $k_1 \in (k_0, 1)$,

$$\begin{aligned} \frac{W(N)_J}{|\Lambda_N(J)|} &\leqslant \mathbb{P}\{F(\omega) \leqslant \alpha_J N \text{ for } \omega \in \Lambda_N(J)\} + 2^{-\alpha_J N} \mathbb{P}\{F(\omega) > \alpha_J N \text{ for } \omega \in \Lambda_N(J)\} \\ &\leqslant 2 \exp(-\min\{\beta_J, \alpha_J \ln 2\} N) \end{aligned} \tag{A.3}$$

since a contact matrix with $M$ free-4-loops is at least $2^M$ times degenerate. Then it follows by proposition 12 that $d(J) \geqslant \min\{\beta_J, \alpha_J \ln 2\} > 0$. □

**Outline of the proof of proposition 12:** Divide $\mathbb{Z}^n$ into disjoint $n$-cubes of edgelength 4. First we remark that at least $aN$ cubes are visited by $\omega \in \Lambda_N(J)$, $a = (1-k_1)/4^n$, and at least $aN/2$ are visited at most by $2/a$ steps. Consider $\Lambda_N^{\alpha N}(J) = \{\omega \in \Lambda_N(J) \text{ s.t. } F(\omega) \leqslant \alpha N\}$, $\alpha \ll 1$. We do two successive operations on $\omega \in \Lambda_N^{\alpha N}(J)$.

(1) We modify the free-4-loops so that the new RW $\widetilde{\omega}$ has $F(\widetilde{\omega}) = 0$. This is obtained by moving the 3rd step to the position of the 1st step of the free-4-loops.

(2) We choose $2\alpha N$ cubes out of the first $aN/2$ visited less than $2/a$ steps. The choice can be made in $\binom{aN/2}{2\alpha N}$ different ways. $\widetilde{\omega}$ passes in a cube not more than $2/a$ times and at each time we replace the path inside the chosen cubes by another one of length increased by 2 which remains on the boundary of the cube and leaving the entry and exit points unchanged. Therefore, the centre of the cubes are now empty. Secondly, we add a free-4-loop in the centre of the cubes the first time that are visited by $\widetilde{\omega}$.

The final RW have length $n \in [N(1+20\alpha), N(1+c_2\alpha)]$ and $I_n/n \leqslant k_1 + c_2\alpha$, $c_2 = 17 + 8/a$. Then using some results of Hamana and Kesten on $\psi(k) = \lim_{N\to\infty} -\frac{1}{N}\ln \mathbb{P}(R_N/N \geqslant k)$ [6], we conclude that, if $\beta_J = 0$, for $\alpha$ small enough the number of constructed RW exceeds the total number of RW with $n \in [N(1+20\alpha), N(1+c_2\alpha)]$ and $I_n/n \leqslant k_1 + c_2\alpha$. Therefore $\beta_J > 0$.

**Outline of the proof of proposition 5:** $\mathbb{P}(R_N/N \leqslant \varepsilon)$ is not exponentially small in $N$ (see, e.g. [6] and proof of (3.1*a*)). Let $J_1 = [0, 1-\varepsilon)$ and $J_2 = [1-\varepsilon, 1]$. Since $d(J_1) > 0$, $W(N)_{J_1}$ is exponentially small compared with $W(N)_{J_2}$ for $N$ large enough. Therefore for large $N$, $W(N) \simeq W(N)_{J_2} \leqslant |\Omega_N|\mathbb{P}(R_N/N \leqslant \varepsilon)$, from which follows (4.14).

The complete proof can be found at http://www-m5.ma.tum.de/pers/ferrari/homepage/download/appendix.ps.gz.

## References


[1] Bahar I, Atilgan A R and Erman B 1997 *Folding Des.* **2** 173
[2] Billingsley P 1965 *Ergodic Theory and Information* (New York: Wiley)
[3] Ferrari P L 2001 Contact matrices for random walks *EPFL-Lausanne Diploma Thesis*
[4] Freed K F 1981 Polymers as self-avoiding walks *Ann. Probab.* **9** 537–56
[5] Haliloglu T, Bahar I and Erman B 1997 Gaussian dynamics of folded proteins *Phys. Rev. Lett.* **79** 3090–3
[6] Hamana Y and Kesten H 2001 A large deviation result for the range of random walk and for the Wiener sausage *Probab. Theory Relat. Fields* **120** 183–208
[7] Havel T F, Crippen G M and Kuntz I D 1979 *Biopolymers* **18** 73
[8] Kesten H 1963 On the number of self-avoiding walks *J. Math. Phys.* **4** 960
[9] Lau K F and Dill K A 1989 *Macromolecules* **22** 3986
[10] Le Gall J-F 1986 *Commun. Math. Phys.* **104** 471–507
Le Gall J-F 1986 *Commun. Math. Phys.* **104** 509–28
[11] Lifson S and Sander C 1979 *Nature* **282** 109
[12] Madras N and Slade G 1996 *The Self-Avoiding Walk Probability and Its Applications* (Cambridge, MA: Birkäuser Boston)
[13] Nieuwenhuizen Th M 1989 Trapping and Lifshitz tails in random media, self-attracting polymers and the number of distinct sites visited: a renormalized instanton approach in three dimensions *Phys. Rev. Lett.* **62** (4) 357–60
[14] Vendruscolo M, Subramanian B, Kanter I, Domany E and Lebowitz J L 1999 Statistical properties of contact maps *Phys. Rev.* E **59** 977–84
[15] Vendruscolo M, Kussell E and Domany E 1997 *Folding Des.* **2** 295
[16] van Wijland F, Caser S and Hilhorst H J 1997 Topology of the support of the two-dimensional random walk *J. Phys.* A **30** 507
[17] van Wijland F and Hilhorst H J 1997 Universal fluctuations in the support of the random walk *J. Stat. Phys.* **89** 119
[18] van Wijland F 1998 Thèse, Université de Paris-Sud, U.F.R. Scientifique d'Orsay