Supporting Information

March 14, 2008

1 Quantities Studied

In this section we shall describe each of the spatial characteristic quantities studied in the main paper: the radius of gyration, average crossing number, miniball radius, convex hull surface area and volume, box dimensions, and thickness.

1.1 Average Crossing Number

Each point on the unit sphere S^2 determines a direction along which one could project the knot to create a two-dimensional shadow of the given configuration. For each of these projection directions $\vec{v}$, one can count the number of times the projected knot crosses itself and call this $cr(\vec{v})$. The *average crossing number* is the average of the integral over the unit sphere of the number of crossings, i.e.

$$\frac{1}{4\pi}\int_{\vec{v}\in S^2} cr(\vec{v})\, dA$$

where A is the surface area. Note that for some directions, the preimage of a crossing point can contain more than two points but that the set of such projection directions is a zero-measure set on the unit sphere. For polygons, the average crossing number can be computed in closed form using ideas proposed by Banchoff [1], implemented by Sullivan in Brakke's Evolver [2], and later recoded for our purposes (see also [3]).

1.2 Radius of Gyration

Let $\overline{v}$ be the center of mass of the vertices of the knot configuration, that is

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i\,.$$

Then the *radius of gyration* is

$$\sqrt{\frac{\sum_{i=1}^{n} |v_i - \overline{v}|^2}{n}}$$

where $|\cdot|$ is the standard Euclidean norm in $\mathbb{R}^3$.

Note that the *average crossing number* and the *radius of gyration* can be thought of as statistical measurements of the shape of the knot in that they are all averages. We now define some geometrical measurements of a different nature.

1.3 Miniball Radius

The *miniball radius* is the smallest radius of a sphere that contains the vertices of the knot and measures the overall size of the knot. We use the implementation of [4] for our calculations.

1.4 Convex Hull Surface Area and Volume

To measure the overall size of a knot, we compute the surface area and volume of the convex hull of the vertex set. The convex hull of a set of points in $\mathbb{R}^3$ is the smallest convex set containing all of the points. Since generically our random knots are non-planar, this convex set is a spatial polyhedron and can be thought of as what one would obtain by shrink-wrapping the set of points. The values were computed using `qhull` [5], an efficient algorithm for computing convex hull volume and surface area as well as other information about the convex hull.

1.5 Box Dimensions

Another way to measure the overall shape of a knot configuration is to find the dimensions of a smallest box that contains the vertex set. We have two techniques for doing so.

The box dimensions were originally defined in [6]. The *box length* is the maximum distance between pairs of vertices of the knot. The chord connecting these vertices is the normal to a family of planes. We project the knot vertices onto one of these planes and the *box width* is the maximum distance between a pair of vertices on the projected knot. We then project the 2-dimensional projected knot onto a line perpendicular to the chord connecting the points that determine the width and define the *box height* as the maximum distance between points on this line. The *box surface area*

and *box volume* are simply the surface area and volume of the box whose length, width, and height are defined above.

1.6 Skinny Box Dimensions

We next define the *skinny box*, another box containing the vertices of the knot. We first search for the shortest distance between two parallel planes that contain the vertices of the knot to define its height. These two planes can be of two types. First, three vertices on the exterior of the convex hull could define one plane with the second plane being parallel and passing through another vertex on the exterior of the convex hull. In the second case, we have two pairs of lines, each through two vertices on the exterior of the convex hull, where the two planes each contain the lines. Since the two planes are parallel and the vertices must be contained between the planes, the two lines must be skew or parallel and, thus, the minimum distance between the two lines is realized on some chord. This chord is the normal direction for each of the planes. The *skinny box height* is the minimum distance between these closest planes that have all of the vertices trapped between them. We then project the knot onto one of the planes and search for the minimum distance between two parallel lines that contain the projected knot. This must be realized by a line through two points on the exterior of the convex hull of the projected knot and a parallel line passing through another point on the exterior of the convex hull. This minimum distance between the two lines defines the *skinny box width*. We then project the 2-dimensional knot onto a line perpendicular to the parallel lines and the maximum distance between the projected points on that line define the *skinny box length*. Again the *skinny box surface area* and *skinny box volume* are simply the surface area and volume of the box whose length, width, and height have just been defined.

Note that in calculating the box dimensions and skinny box dimensions, the implicated vertices must lie on the exterior of the convex hull. We use `qhull` [5] in 2 and 3-dimensions to speed the calculations.

1.7 Thickness

One measure for the compaction of a polygon is radius of a thickest non self-intersecting tube that can be placed about the polygon. Roughly speaking, the *thickness radius* $R(K)$ measures the maximum radius of this tube. The thickness

radius is the minimum of two functions, $MinRad$, which measures the local curving of the polygon, and half of the *doubly-critical self-distance*, which measures the minimum direct distance between arc-wise distant portions of the polygon. Details for the definition of thickness can be found in [7, 8, 9]. The well-studied *ropelength* is a measure of knot compaction and is the ratio $Length(K)/R(K)$. The knot configurations that minimize the ropelength are sometimes called *ideal knots* and have been shown to predict some of the physical and statistical properties of random knots (see e.g. [10, 11, 12, 13, 14]).

2 Monte Carlo Markov Chain Method

In this section, we present our Markov Chain Monte Carlo (MCMC) algorithm for fitting functions to the data and computing error bars for the equilibrium length values. Our reason for using MCMC versus non-linear regression is to facilitate the computation of error bounds for the crossing between two fitting curves. With the MCMC approach, this error computation is straightforward, albeit CPU-intensive.

For a given knot type and spatial quantity, we will have a list of sample data $\vec{y_n}$ for each of the different numbers of edges n. Recall that the $\vec{y_n}$ are the results from completing the averaging procedure (explained in the main text) so that the data tests as being normally distributed at the 95% confidence level for each knot type/spatial quantity pairing. The distributions, with the exception of the thickness, are lognormal at a given number of edges (see the following section).

To estimate the parameters A, B, C of the mean function $g(n) = n^{d\nu}(A + B/\sqrt{n} + C/n)$, we implement a MCMC algorithm known as Metropolis sampling [15, 16]. This algorithm requires the following assumptions about our data:

- All values in $\vec{y_n}$ (for a given knot type and spatial quantity) observed at a given n-value (a fixed number of edges) are independent; $\vec{y_n}$-values are also independent across n values;
- All values in $\vec{y_n}$ at a given n-value come from a known density function f (in our case, f is a normal distribution);
- The mean of the distribution f (our spatial values) at a given n-value is $g(n)$;

- The unknown parameters A, B, and C are treated a priori as random quantities, and hence are assigned (prior) probability distributions.

For our application, assumption 1 is easy to accept, as independence is inherent in the data generating process. The averaging procedure does not affect the independence of the data points.

Assumption 2 is mild in this case. As described in the main text, for each knot type and spatial quantity, over 95% of the distributions pass the Shapiro-Wilk test at 95% confidence levels.

Assumption 3 is a bit tenuous, but ultimately not a problem. For the average crossing number, the fitting functions have been derived in [17]. We will reserve our discussion of thickness until the end of this explanation.

For the remainder of the quantities, the correction term $A+B/\sqrt{n}+C/n$ comes from [18], where it was originally applied to the scaling of the radius of gyration for self-avoiding walks. Scaling functions for self-avoiding polygons are typically used also for modeling polymers with topological constraints [19]. The asymptotic term $n^{d\nu}$ is the generalization of the scaling function n^{ν} used for the radius of gyration. While we have no formal proof that the fitting function $n^{d\nu}(A + B/\sqrt{n} + C/n)$ extends to the quantities explored in this paper, we have three reasons why we believe this is a reasonable assumption.

First, one would expect linear sorts of quantities, such as miniball radius and radius of gyration, to scale similarly, and the higher dimensional quantities to scale analogously. Second, even if the correction term should be different, our results here will be minimally affected. Here is our explanation of why. Assume that the correction term instead looks like $(A' + B'/n^{\alpha} + C'/n^{\beta})$ where α and β are unknown. In our Markov chains, the values of B and C are intimately related. By increasing B, one can decrease the value of C and get a similar looking curve, and we see this behavior repeatedly in our computations. In fact, the value of A affects the curve much more than B and C. Namely, while A affects the asymptotic scaling, the values of B and C account only for the "curviness" of the fit in the small-edge regime and ultimately die off in the asymptotics. By adding varying α and β, we would only get more inter-relations (now between all of α, β, A', and B'), but the resulting "curviness", and thus fit, in the small-edge regime will be minimally different from fixing the values of $\alpha = 1/2$ and $\beta = 1$. This is, in fact,

one of the strengths of this type of analysis. Third, the fits are quite convincing. Ultimately, we simply need a correction term with some curving and the choice here provides that.

For thickness, our choice of fitting function came from our efforts to find an analogous function. The log-log plot is well fit by the function $A+B/\sqrt{n}+C/n$, which results in a fitting function of $\exp(A + B/\sqrt{\ln(n)} + C/\ln(n))$.

Assumption 4 is a matter of modeling preference, and allows for the use of MCMC algorithms to update model parameters (standard Bayesian statistical methodology). Here we choose independent uniform (non-informative) prior distributions on A, B, and C over sufficiently wide intervals. Bayes' Theorem combines these uniform priors with the normal data model to produce the (joint) posterior probability distribution of the parameters A, B, and C. Using $\pi(A, B, C|\vec{y}, \vec{x})$ to denote this posterior, we have:

$$\pi(A, B, C|\vec{y}, \vec{x}) \propto \prod_i \prod_j f(y_{ij}) \ .$$

After initializing parameters A, B, and C, the Metropolis algorithm proposes a new value of, say, the parameter A and then accepts this proposed value with probability α:

$$\alpha = min \left\{ 1, \frac{\pi(A^*, B, C|\vec{y}, \vec{x})}{\pi(A, B, C|\vec{y}, \vec{x})} \right\} \ ,$$

where A^* is the proposed value of A. Repeating this process for all three parameters thousands of times yields thousands of realizations from the joint posterior π. The realization $\{A, B, C\}$ at any iteration can be used to draw the mean function g; thousands of such (independent) realizations yield thousands of mean functions g that are likely to have produced the data.

We use `gnuplot` [20] to determine initial values of A, B, and C to minimize the initial burn-in time (used to reach a reasonable fit). The most delicate part of the computations is determining appropriate distributions for A, B, and C with the hopes of quickly eliminating the autocorrelation in the Markov Chain of parameter values. When a new parameter value, say A^*, is proposed, it is chosen from a normal distribution whose mean is the current value of A and variance is manually chosen. For small chosen variances, the posteriors will be almost the same, and thus the α value will be close to one. This results in A^* being accepted at nearly every step, but with small changes in A values between steps. On the other hand, if the variance is chosen to be large, then few

A^* values are accepted but the changes that are accepted are significant. For a Monte Carlo chain, we do not want to analyze every step of the walk (i.e. the curve determined by A, B, and C at each repetition) because of the inherent autocorrelation between the parameter values in a single step. Thus, we compute the autocorrelation function and only consider parameter values after the autocorrelation is below 5%. In other words, when the variance is chosen too low, the values change often but by small amounts resulting in long lag times. When the variance is chosen too high, then the parameter values almost never change, and again we have long lag times to remove the autocorrelation.

For this effort, we chose the variance as a percentage of the initial parameter values. The percentage was different for each parameter value, and was chosen after much experimentation. Ultimately, we were able to find percentages which resulted in lags of around 300 (i.e. in all three parameter values losing the autocorrelation in 300 steps of the algorithm). Thus, we required 300 steps of the algorithm to determine one proposed fitting curve.

As described previously, we compute 1000 likely fitting curves for each pairing of knot type (in addition to the phantom polygons) and spatial characteristic. We then compute 1000 intersection points from these curves as follows. In the case of the 3_1 knot with the miniball radius, we have 1000 likely fitting curves. For the phantom polygons with miniball radius, we also have 1000 likely fitting curves. We compute the intersection between the first curve for the 3_1 knots versus the first curve for the phantom polygons. We continue with the second curves, third curves, etc. This generates a list of 1000 likely crossing points. The smallest 2.5% and largest 2.5% are removed, and what remains defines the 95% confidence interval for the equilibrium length of the 3_1 knot with respect to the miniball radius. This procedure is then repeated for each of the knot types in combination with each of the spatial quantities. In Table 1 of the main paper, the equilibrium length values and errors have been rounded to the nearest integer.

3 Distribution of Values at a Fixed Number of Edges

In this section, we present evidence that the distributions of the radius of gyration, miniball radius, box dimensions, and

convex hull surface area and volume are lognormal. Note that recognizing the actual distribution of the samples at a given number of edges is not required when implementing the MCMC grouping procedure described in the main text. For the sake of simplicity, we have chosen the phantom polygons and 3_1 knots with 200 edges for our demonstration. We computed histograms of the raw data and the log of the data for each of the spatial quantities (with the exception of thickness). Each of these was fit with a normal curve with the same mean and variance as the data. The graphs clearly show that the raw data is not fill well by a normal distribution, but the log of the distributions are well-fit by a normal distribution, i.e. the distributions are lognormal. The remainder of this section shows these plots for the 3_1 knots and phantom polygons with the representative quantities: average crossing number, radius of gyration, miniball radius, box surface area, and convex hull volume.

References

[1] Banchoff, T., *Indiana Univ. Math. J.* **1976**, *25*, 1171.

[2] Brakke, K.; Sullivan, J., Surface Evolver, http://www.geom.uiuc.edu/software/download/, Program for visualizing and energy minimizing knots.

[3] Cimasoni, D., *J. Knot Theory Ramifications* **2001**, *10*, 387.

[4] Miniball, http://miniball.sourceforge.net/, Program for computing smallest enclosing ball.

[5] Qhull, http://www.qhull.org, Program for computing the convex hull.

[6] Millett, K. C.; Rawdon, E. J., *J. Comput. Phys.* **2003**, *186*, 426.

[7] Rawdon, E. J. 1997, *The Thickness of Polygonal Knots*, Ph.D. thesis, University of Iowa.

[8] Rawdon, E. J. 1998, in *Ideal knots*, pp. 143–150, World Sci. Publishing, Singapore.

[9] Rawdon, E. J., *J. Knot Theory Ramifications* **2000**, *9*, 113.

[10] Cerf, C.; Stasiak, A., *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3795.

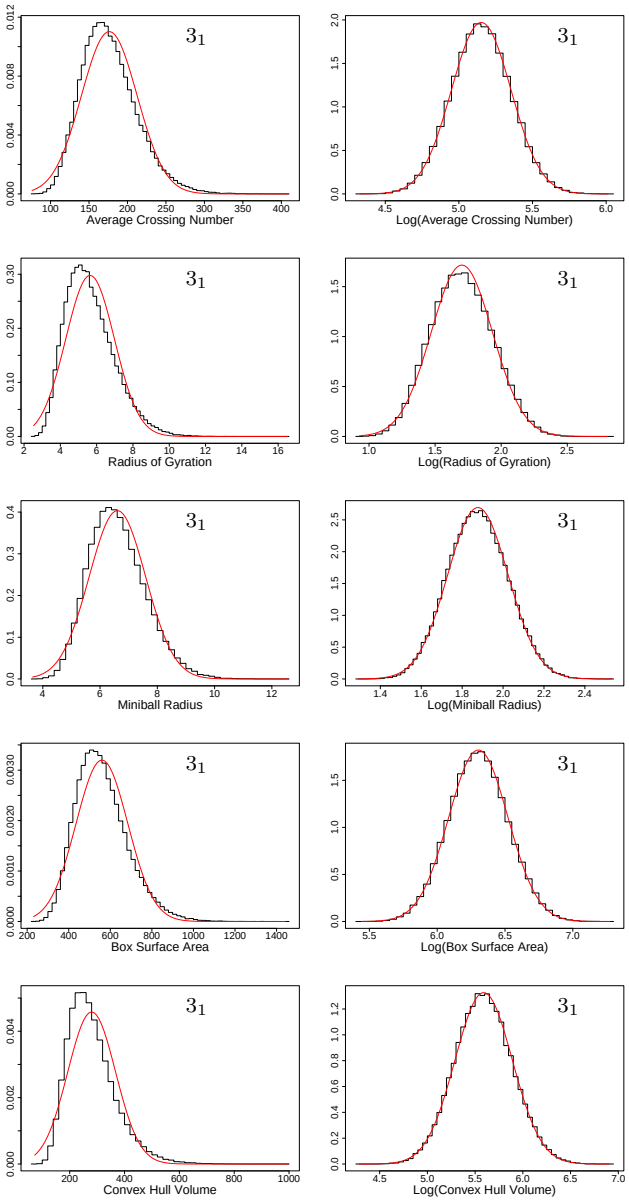

Figure 1: The distributions for several different quantities measured on 3_1 knots with 200 edges appear to be lognormal. In the left column, normalized histograms for the data is compared to normal distributions with the same mean and variance. In the right column, normalized histograms for the log of the data is compared to normal distributions with the same mean and variance.

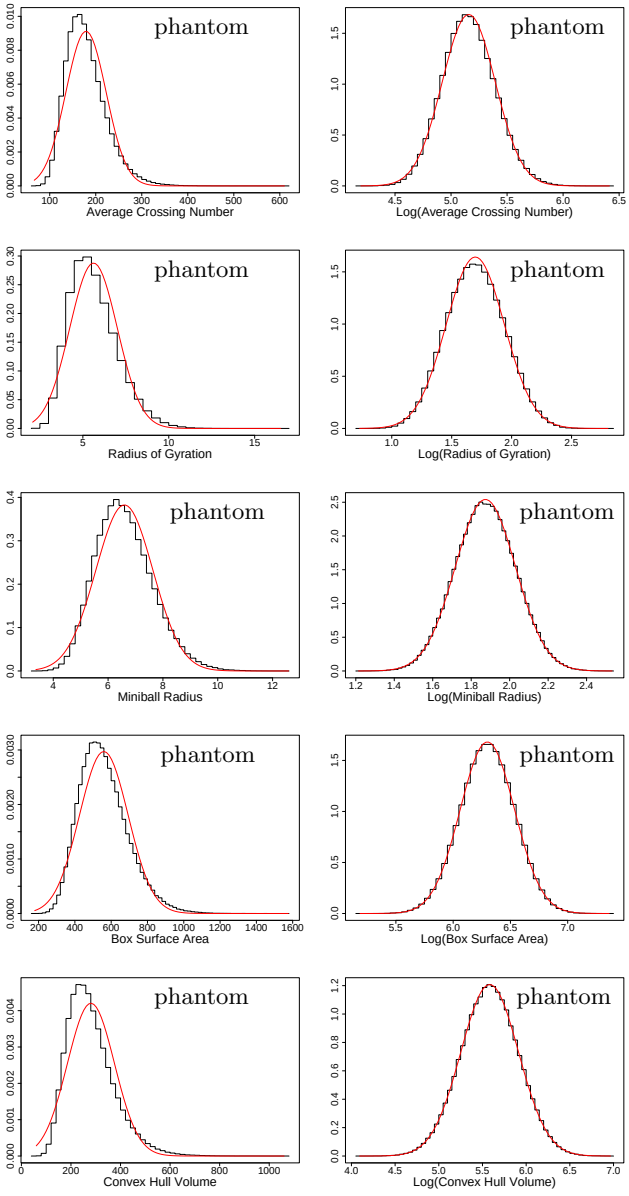

Figure 2: The distributions for several different quantities measured on phantom polygons with 200 edges appear to be lognormal. In the left column, normalized histograms for the data is compared to normal distributions with the same mean and variance. In the right column, normalized histograms for the log of the data is compared to normal distributions with the same mean and variance.

[11] Grosberg, A. Y. 1998, in *Ideal knots*, vol. 19 of *Ser. Knots Everything*, pp. 129–142, World Sci. Publishing, Singapore.

[12] Maritan, A.; Micheletti, C.; Trovato, A.; Banavar, J. R., *Nature* **2000**, *406*, 287.

[13] Stasiak, A.; Katritch, V.; Bednar, J.; Michoud, D.; Dubochet, J., *Nature* **1996**, *384*, 122.

[14] Stasiak, A.; Dubochet, J.; Katritch, V.; Pieranski, P. 1998, in *Ideal knots*, pp. 1–19, World Sci. Publishing, Singapore.

[15] Metropolis, N.; Rosenbluth, A.; Rosenbluth, M.; Teller, A.; Teller, E., *The Journal of Chemical Physics* **1953**, *21*, 1087.

[16] W. R. Gilks, S. Richardson, D. J. S. (ed.), *Markov chain Monte Carlo in practice*, Chapman and Hall, London 1996.

[17] Diao, Y.; Dobay, A.; Kusner, R. B.; Millett, K. C.; Stasiak, A., *J. Phys. A* **2003**, *36*, 11561.

[18] Orlandini, E.; Tesi, M. C.; Whittington, S. G.; Sumners, D. W.; Janse van Rensburg, E. J., *J. Phys. A* **1994**, *27*, L333.

[19] Deutsch, J. M., *Phys. Rev. E* **1999**, *59*, 2539.

[20] gnuplot, http://www.gnuplot.info, Program for plotting and fitting data.