Analysis of the Hydrogen Bond Network of Water Using the TIP4P/2005 Model

D.C. Elton
daniel.elton@stonybrook.edu
Stony Brook University AMS 591 Final Project

December 18, 2012

1 Introduction

Liquid water is one of the most fascinating substances in the universe. It possesses a number of anomalous properties which cannot be found in other liquids - on his website, Martin Chaplin identifies 69 different anomalous properties.[4] Chief among these are anomalously high melting and boiling points relative to water's small molecular weight, water's expansion of volume upon freezing, and the lowering of the freezing point with pressure. Closely related to these is a class of response function anomalies:

- The isothermal compressibility K_T has a minimum at 46 C and then increases at lower temperatures. (Usually K_T decreases monotonically with T.)
- The specific heat C_P has a minimum at 36 C and increases at lower temperatures. (Usually C_P decreases monotonically with T.)
- The thermal conductivity κ of water is unusually high and increases with temperature until reaching a maximum at 130 C. (Usually κ decreases monotonically with temperature)

All of the anomalous properties of water are connected to the fact that water forms an extensive hydrogen-bond network. Water is the only low molecular-mass molecule to form a hydrogen bonded liquid (other examples of hydrogen bonded liquids are ammonia, which forms only weak hydrogen bonds, and chain alcohols). It should be noted that, despite the fact there are only a few hydrogen bonded liquids, hydrogen bonds are found in many other contexts such as in proteins, DNA and "hydrogen bonded solids" such as cellulose, nylon and other polymers.

The hydrogen bond network of water has been studied extensively over the past one hundred years using a wide variety of experimental and theoretical techniques. Among these has been network analysis, which is what this paper will focus on.

In this paper at times we will use the language of network science. Nodes correspond to water molecules and links will correspond to hydrogen bonds. The degree of a molecule refers to the number of molecules a given molecule is connected to, and the max number of possible links a molecule can make is called the "coordination" number of the molecule. For water the coordination number is four. Each water molecule can accept two hydrogen bonds and can also donate two. The hydrogen bond in water is not symmetric, ie. the proton (hydrogen) involved in the bond remains closer to the donor molecule than the acceptor. Thus, one could create a directed network by defining a direction from the donor to the acceptor. However we have not had any reason for doing this so far so all of our networks are undirected.

2 Previous studies

The first study of the H-bond network which used techniques belonging to network science was published by Rahman & Stillinger in 1973.[16] Their simulations were done with the "ST2" model of water[26] which is an early five site highly tetrahedral model which was later supplanted by more carefully fitted models. The overall accuracy of the ST2 model is questionable – for instance

the observed density minimum of the model is too high (300 K vs. 277 K) and the self diffusion constant was described by its creators as "highly erroneous". Still, some interest in the ST2 model continues to this day, especially after it was shown to exhibit a liquid-liquid phase transition at low temperatures.[10] In addition to their use of the ST2 model, Rahman & Stillinger's definition of a hydrogen bond is also questionable. They a simple definition which is illustrated in figure 1. Two molecules are considered to be bonded if the potential energy of interaction is below some cutoff V_{HB}.

In this early work, Rahman & Stillinger studied the number of non-short-circuited polygons (NSCPs) of size j found in the H-bond network of 216 ST2 molecules at 283K. A polygon (network term: cycle) is considered not to be "short circuited" if there are no links between nodes going through the interior of the polygon. In other words, two polygons (such as two triangles) may share an edge, but the larger polygon which encompasses both should not be counted. The distribution of NSCPs is of interest because it allows one to compare the liquid hydrogen bond network with the hydrogen bond network found in solid water. In solid water, the allowed sizes for NSCPs is fixed by the crystal structure. Table 2 shows the sizes of NSCPs for different phases of ice.

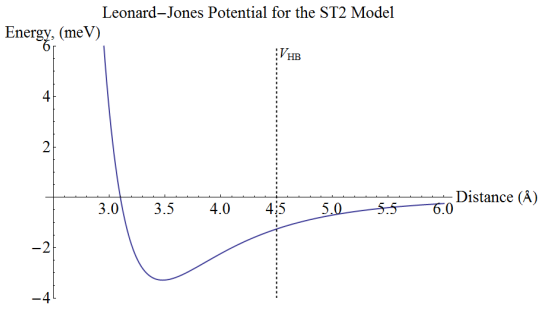

Figure 1: Schematic of the H-bond definition used by Rahman & Stillinger. (The actual potential also includes electrostatic interactions)

Rahmann & Stillinger proceeded to calculate the distribution of NSCPs of size j for four different choices of V_{HB} between -2.121 and -4.848 kcal/mol. They found that their choice of $V_{HB} = -4.848$ kcal/mol was too restrictive and gave a distribution of j which was limited to $j = 5$ and $j = 6$ For their other three choices of V_{HB} they found distributions which were peaked around $j = 6$ but which also had sizable tails extending out to $j = 11$, which was the highest j they considered. From the distributions it was clear that NSCPs with j greater than 11 were certainly not rare. These findings cast serious doubt on earlier ideas that the hydrogen bond network of water might be considered (at least locally) very similar to that of ice, yet "disrupted" enough to make water a liquid.

Table 1: Sizes of non short-circuited polygons in various phases of ice. Table copied from (Rahman, 1973), water ice data was originaly taken from (Einsenberg, 1969).

Phase	Allowed NSCP Sizes
Ice Ih	6
Ice Ic	6
Ice II	6
Ice III	5
Ice V	4
Ice VI	4,8
Ice VII	6
Ice VIII	6
Cl_2 /CH_4 / Xe hydrates	5,6

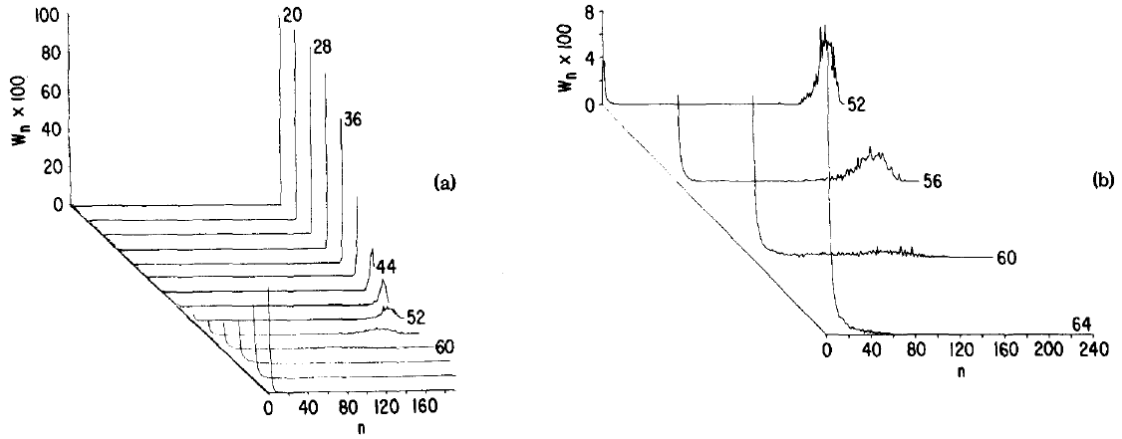

Figure 2: Figure from [6]. (a) The mole precentage of water molecules participating in clusters of size n for various values of $V_{HB} = -k\epsilon$. (b) Same data as in (a) but magnified to show the percolation transition near $k = 60$. The number of molecules in the simulation was 216, so values of n near 200 correspond to clusters that fill all of space.

Geiger, Stillinger & Rahman continued studying the hydrogen bond network of water in 1978[6] using trajectories from earlier simulations of the ST2 potential.[26][27] In this work they looked at the distribution of cluster sizes for various choices of V_{HB}. V_{HB} was parameterized by an integer k such that $V_{HB} = -k\epsilon$ where ϵ =.07575 kcal/mol. These distributions are shown in figure 1. For high values of k, the choice of V_{HB} was too restrictive, so the hydrogen bond density was low and the cluster size distribution decayed rapidly. At $k = 60$ a remarkable percolation transition was observed to occur. In the language of network science, at $k \approx 60$ the percolation threshold is reached and a giant component is formed. The giant component extends through the entire simulation cell, although there are molecules which are not bound to it. For $k = 24$ and smaller, the criteria becomes too free and all molecules are "bound" into a single cluster. They then went on to analyze $\bar{n}_{HB}$, the average number of hydrogen bonds per molecule. There is no easy experimental way to measure n_{HB} and prior to computer simulations only various simplified theoretical models could be used. These models gave a wide range of predictions for $\bar{n}_{HB}$, from as much as $\bar{n}_{HB} = 3.9$ to as little as $\bar{n}_{HB} = 1.84$[6]. They were then able to plot the average cluster size $\bar{j}$ vs. $\bar{n}_{HB}$. Their most dramatic result came from thier simulation with $N = 1728$ molecules, where they observed that $\bar{j}$ shot up to close to the size of they system when $\bar{n}_{HB} \approx 1.3$. For their two other simulations with $N = 216$ the sudden increase in $\bar{j}$ occurred when $\bar{n}_{HB} = 1.5 - 2$.

The results of Geiger & Stillinger are easily understood as a type of percolation transition. In a percolation transition, the order parameter is the concentration of bonds x (or equivalently, $\bar{n}_{HB}$). When $x \ll x_c$ only small isolated clusters appear. For an infinite system, the mean cluster size $\bar{j}$ diverges at $x = x_c$. Above x_c the system is dominated by one cluster of infinite size (the "giant component") with the possibility of many small disconnected clusters scattered throughout. Above x_c any two nodes (molecules) are likely to be connected through some path in the giant component and the chance that one of the two molecules is not in the giant component becomes very small. In a finite system, something similar happens although now the size of the giant component is limited to the size of the system.

The observation that the percolation transition occurred when $\bar{n}_{HB} \approx 1.3$ confirms a theory of Stockmayer, which drew on the earlier "gellation model" of Flory.[28] Stockmayer uses an order parameter called the "degree of polymerization" α which is the fraction of possible bonds formed by each molecule. The maximum number of a bonds a molecule can form he calls the "functionality" f. The percolation phase transition occurs when $\alpha = \alpha_c$ where α_c is given by:

$$\alpha_c = \frac{1}{f-1} \tag{1}$$

For water the obvious choice is to have $f = 4$, yielding $\alpha_c = 1/3$. The average number of hydrogen bonds at the critical point is then given by $\bar{n}_{HB} = f\alpha_c = 4/3 \approx 1.3$.

2.1 The percolation model of Stanley & Teixeria

It appears that the next major study of the H-bond network was done by Stanley & Teixeira in 1980 when they looked in more detail at the percolation transition.[25] Stanley & Teixeira's analysis was quite comprehensive, with additional Monte Carlo simulations and calculations of cluster sizes given in an 1984 follow up paper.[3] It is worth noting that network connectivity information is impartial to the spatial positions of molecules and that percolation occurs in "connectivity space" and not in the real space which physicists are used to thinking in terms of. In the cases we are interested there is clearly a link between the connectivity and real spaces since molecules which are connected will be spatially close. To help visualize the connectivity space and map it onto real space, percolation is often considered to happen on a lattice - although this is by no means necessary – for instance Stockmayer's theory does not assume a lattice. Percolation is most easily studied on a square lattice (at least in the opinion of this author), but for water, the tetrahedral lattice makes more sense. Somewhat counter-intuitively, despite the fact that both the square lattice and the tetrahedral lattice have coordination of four ($z = f = 4$), the percolation thresholds for them are different. In percolation, the topological properties of the lattice are very important. Table 2.1 shows percolation thresholds for bond percolation and site percolation for various lattices.

	lattice	bond	site
1D	any	1.0	1.0
2D	square	1/2	.592746
	triangular	.34729	1/2
	honeycomb	.65271	.6962
3D	simple cubic	.2488	.3166
	BCC	.1803	.246
	FCC	0.1992365(10)	0.1201635(10)
	HCP	0.1992555(10)	0.1201640(10)
	tetrahedral(ice)	0.388(10)	0.433(11)

Table 2: Bond percolation and site percolation thresholds for some well known lattices. These numbers are concentrations of bonds/sites and are equivalent to probabilities. In addition to the exact results of 1/2 for the square bond percolation and triangular site percolation, two other exact results are known - the triangular bond threshold ($2\sin(\pi/18)$) and the honeycomb bond threshold ($1\text{-}2\sin(\pi/18)$). The rest of the values are extrapolated from computer simulations.

Stanley & Teixeira choose to focus their attention on four-bonded molecules, working on a cubic lattice. If the probability of a bond is $p_B = \bar{n}_{HB}/4$ then the probability that a molecule is four bonded is simply p_B^4. Next, they demonstrate the somewhat surprising result that four-bonded molecules tend to clump together, despite the fact that bonds are independent and uncorrelated. The reason for this is simply do to the combinatorics of the situation.

The fact that the four bonded molecules are correlated can be proven mathematically as follows. To avoid confusion, Stanley refers to four bonded sites as "black dots" because that is how he labeled them in his diagrams (see fig 3). The probability that a molecule is a "black dot" is simply $c = p_B^4$. The probability that a molecule is four-bonded molecule and surrounded by neighbors which are all four-bonded is:

$$P = p_B^4 * p_B^3 = c^{7/4} \tag{2}$$

The probability that a molecule is four-bonded and surrounded by neighbors which are *not* four-bonded is:

$$N = p_B^4(1 - p_B^3)^4 = c(1 - c^{3/4})^4. \tag{3}$$

Now consider if the black dots were distributed randomly with probability c (the uncorrelated case). Then the probability that a site is a black dot and surrounded by sites which are *not* black dots is:

$$N^* = c(1 - c)^4 \tag{4}$$

Likewise the probability that a molecule is four-bonded and surrounded by molecules which are four-bonded is:

$$P^* = c^2 \tag{5}$$

We find that $N < N^*$ or that $P > P^*$ which both imply that the distribution of black dots (four-bonded molecules) is correlated - ie. they clump together.

We can gain more insight comparing by looking at $P/P*$ and $N/N*$ as a function of p_B

(designated p here for simplicity):

$$\frac{P}{P^*} = \frac{c^{7/4}}{c^2} = \frac{p^7}{p^8} = \frac{1}{p}$$
$$\frac{N}{N^*} = \frac{c(1-c^{3/4})^4}{c(1-c)^4} = \frac{(1-p^3)^4}{1-p^4)^4} \tag{6}$$

At low temperatures, we expect $p \to 1$, whereas at high temperatures we expect $p \to 0$. The behaviors of P and N in the correlated and uncorrelated cases are shown more explicitely in figures 4 and 5.

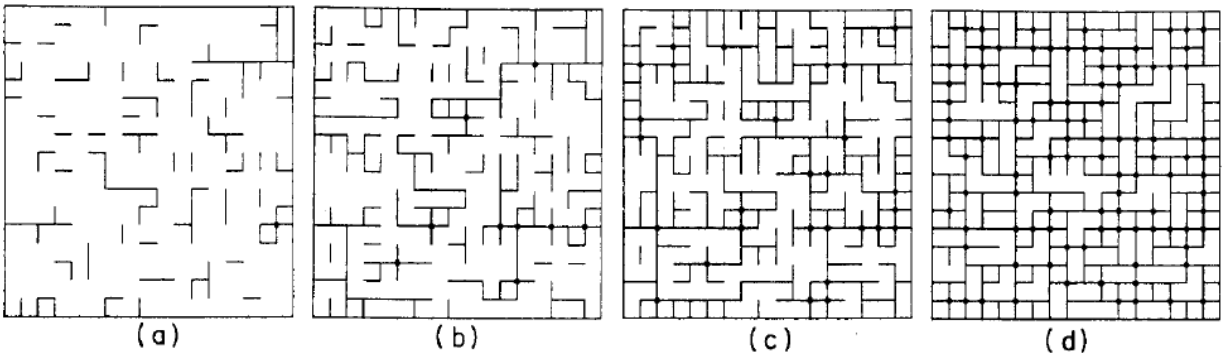

Figure 3: Bond percolation example from [25]. (a) $p_B = .2$ (b) $p_B = .4$ (c) $p_B = .6$ (d) $p_B = .8$. Here these simulations are done on a 2D square lattice, whereas for water a 3D tetrahedral lattice is more appropriate. The four-bonded molecules are shown as black dots. The bond percolation threshold is 0.5 but the percolation threshold for four-bonded molecules is .56. So, even at $p_B = .8$ the percolation threshold for four-bonded molecules has not been reached, since $f_4 = p_B^4 = .4096$.

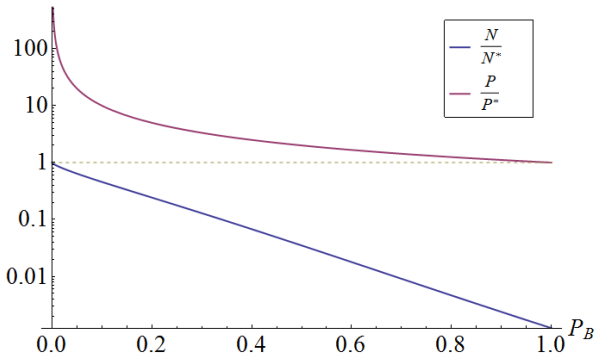

Figure 4: N/N^* and P/P^* vs p_B. At low temperatures, $p_B \to 1$. P corresponds to the number of four-bonded molecules which are surrounded by four-bonded molecules in the correlated (no star) and uncorrelated (starred) cases. N is the probability of finding a four bonded molecule which is not surrounded by four bonded molecules. The interesting thing here is that at low temperatures P is the same in both the correlated and uncorrelated case, but at high temperatures (small p_B), P becomes exponentially larger in the correlated case (of course, at the same time, $P \to 0$ at high T).

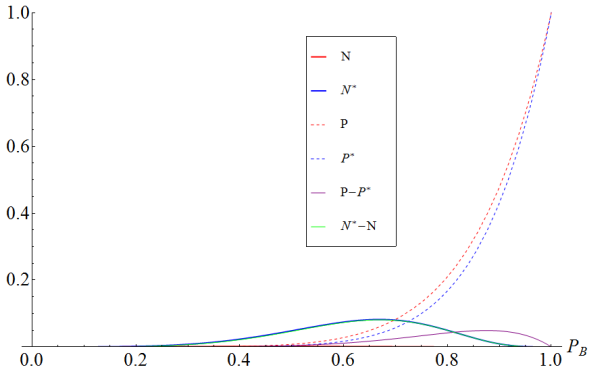

Figure 5: Graphs of P, P^*, N and N^*. N is so small that it appears near zero through the entire range on this graph. It is interesting to see how the difference changes with p_B, which goes to 1 as $T \to 0$.

The end result of this is that there are unavoidable correlations in the distribution of four-bonded sites. Stanley & Teixera assume that the bondedness of a molecule is proportional to the amount of volume that it takes up:

$$V_0 \lesssim V_1 \lesssim V_2 \lesssim V_3 \lesssim V_4 \tag{7}$$

Thus four-bonded clusters will have lower density, leading to density fluctuations in the liquid. The isothermal compressibility is directly related to density fluctuations, according to linear response theory:

$$\begin{aligned} K_T &\equiv \frac{1}{\rho}\left(\frac{\partial \rho}{\partial P}\right)_T \\ K_T &= \frac{V}{k_B T}\frac{\langle \rho - \langle \rho \rangle \rangle^2}{\langle \rho \rangle^2} = \frac{V}{k_B T}\frac{\langle \rho^2 \rangle - \langle \rho \rangle^2}{\langle \rho \rangle^2} \end{aligned} \tag{8}$$

Normally K_T decreases monotonically with lower temperature. However, the density fluctuations from four-bonded site correlations will create an anomalous effect – by increasing K_T at lower temperatures. Using a variety of experimental evidences, Stanley & Teixeira argue that this is what is responsible for the observed minimum in K_T at 46 C.

In addition to creating clusters of lower density, four-bonded sites will also create clusters with lower entropy. This yields to anomalous behavior in the specific heat at constant pressure C_P. Using similar arguments, Stanley & Teixeira show how the four-bonded clustering can explain two other static response functions – the constant volume specific heat C_V and the thermal conductivity.

3 Effect on dynamic response functions - in particular the dielectric response $\varepsilon(\omega, T)$

The dynamic response functions which are usually of interest are step response functions. For instance, if an electric field is applied and suddenly turned off, how does the polarization $P(t)$ decay with time? A related question is, at a molecular scale, if a molecule has an orientation $\vec{V}$ at time t_0, how does its orientation decay with time, on average? Usually to fairly good accuracy such decays can be described as single exponential decays with a time constant τ. In their work, Stanley & Teixeira consider the rotational decay time constant τ_R for a single molecule. They make a number of gross simplifications – first they assume that molecules forming 1, 2, or 3 bonds are "immobile". The fraction of immobile molecules is denoted F_I

$$F_I \equiv f_2 + f_3 + f_4 \tag{9}$$

They next assume that hydrogen bonds are broken on a timescale τ_{HB} and that snapshots of water which are separated by times greater than τ_{HB} are uncorrelated. This is a drastic assumption, since "memory effects" surely extend beyond the average hydrogen bond lifetime τ_{HB}. Now consider l snapshots, each separated in time by τ_{HB}. Then the probability of a molecule remaining immobile for l snapshots in a row is F_I^l. Now we wish to find the time $\tau_R = l\tau_{HB}$ required for

the probability of being immobile to decrease by a factor of $1/e$ (in their paper, Stanly & Teixeira used $1/2$, but $1/e$ is a more standard measure of decay). In other words:

$$F^{\tau_R/\tau_{HB}} = \frac{1}{e}\tau_R = \frac{\tau_{HB}\ln(\frac{1}{e})}{\ln(F_I)} \tag{10}$$

Stanley & Teixera also proposed an approximate temperature dependence for p_B:

$$p_B = 1.845 - .004T \tag{11}$$

Using the relation $F_I = p_B^2 + p_B^3 + p_B^4$ and equation 11 one can derive an expression for $\tau_R(T)$.

One can relate τ_R to the diffusion constant since experimentally it is observed that $\tau_R D =$ constant. Using experimental data for τ_R one can also use these relations to estimate p_B and/or τ_{HB}. This was done by Nabokov & Lubimov in 1988.[13] The microscopic relaxation time τ_R is hard to access experimentally, so they used the macroscopic (Debye) relaxation time τ_D and a (rather questionable) theoretical relation between τ_D and τ_R.[13]

3.1 Subsequent work using network analysis

In a 1984 paper, Speedy argued that pentagons may be a better way of identifying low density patches than four-bonded molecules.[23] Work by Speedy & Mezei in 1985 studied the concentration of pentagons and paired pentagons using the ST2 and MCY models.[24] They found that the concentration of pentagons has a strong temperature dependence and they even calculated radial distribution functions (RDFs) for pentagons and paired pentagons.

In 1979 Sceats & Rice proposed a "zeroth order random network model of liquid water"[21] which proposed modeling water as a completely connected, highly tetrahedral network. Between 1979 and 1981 they published a number of papers on what they called the Random Network Model (RNM) of water.[18][19][20][17] Further work by Belch & Rice looked at the distribution of NSCPs (which they call "rings") in TIPS-2 water at five temperatures between 243 and 313 K.[2] They found that rings of size six (hexagons) were the most common ($\approx$ 20% of molecules belonged to hexagons at 298 K vs 15% in pentagons, 7.6% in decagons, .3% in triangles and 2.4% in no ring).

A study in 1995 analyzed NSCPs in ST4 water and compared bulk water with hydration water around small hydrophobic solutes such as methane. It was found that the cage structures around hydrophobic solutes contain a lot of pentagons, wheras in bulk water much larger polygons dominate.[7]

Work was published in 1990 which did a network analysis of the SPC/E model.[12] In 1996 Shiratani & Sasai proposed the "local structure index" (LSI), which is a continuous parameter that measures the variance of the radii of molecules within a sphere $r < 3.7\mathring{A}$ around a given molecule.[22] The LSI measure is not really related to network analysis, but their analysis essentially reiterates the ideas of Stanley & Teixeira about four-bonded patches, but with a discrete bondedness measure being replaced by LSI. LSI is directly related to bondedness because when molecules are four bonded they have a large LSI whereas when they have no bonds they will have an LSI close to zero. Shiratani & Sasai then go on to extend the work of Stanley & Teixeira by looking at the fluctuations in LSI over time and computing the power spectra of LSI(t).

There has also been some work done investigating the percolation threshold in superheated and supercritical water.[15][5] At slightly below the boiling temperature, the hydrogen bond network of water is still very well connected and water is not near the percolation threshold (this is seen very clearly in the 400 K simulations reported below). Thus boiling/condensation and crossing the percolation threshold from above or below are not related as one might naively expect. However, there has been some work looking at water beyond the critical point of the phase diagram.[15] SPC/E Monte Carlo simulations show that if one extends the vapor-liquid coexistence curve beyond the second-order critical point at (647 K, 22 MPa) then this extension effectively becomes a "percolation line" - water above the percolation line (higher pressure) has a giant component whereas water below does not.[15] Another paper hypothesizes that there is a maximal limit to superheating which corresponds to the percolation threshold.[5] As with any percolation transition, at the percolation line there is a great deal of self-similar structures and the distribution of cluster sizes is described by a power law.

4 Methods

4.1 Identification of hydrogen bonds

The nature of the hydrogen bond has been studied in great detail. It is not our task here to describe how hydrogen bonding works, rather we simply wish to know how to identify a hydrogen bond.

IUPAC, the international standards organization for chemists, laid out a set of six criteria which should be satisfied by a hydrogen bond in 2011.[1] Although six criteria sounds like a lot, these criteria are left very general, with two of the more restrictive criteria being that the Gibbs energy of formation of the hydrogen bond be greater than the thermal energy of the system and that there be partial charge transfer between the donor and acceptor leading to partial covalent bond formation within the hydrogen bond. In a classical molecular dynamics simulation we cannot see charge transfer, and in fact all that one does see is the coordinates of the atoms. Thus we need a criterion in terms of the coordinates (geometry) of the situation, which is something IUPAC does not provide. A traditional geometric criteria is that the oxygen-oxygen distance be less than 3.5 $\mathring{A}$ and that the bond angle be less than 35°. 3.5 $\mathring{A}$ is just slightly more than the distance to the first minima in the oxygen-oxygen radial distribution function (at ≈ 3.2 $\mathring{A}$). This criteria undoubtedly leaves out some weaker bonds which would still fall under the IUPAC criteria, and may at times also over count in the rare instance that a third water molecule is in the proximity. Ideally, one should test the dependence of the network properties on the definition.

Sometimes, the max degree of a molecule is strictly limited to four in the H-bond definition since there are four binding sites. In the event that a molecule has more than four hydrogen bonds, only the strongest four are counted. However, degrees of five are also accepted by chemists as physically reasonable if one binding site has a bifurcated hydrogen bond. A bifurcated hydrogen bond is thought to form in water during the short period of time between the breaking of one bond and the forming of another. If one accepts bifurcated hydrogen bonds, then even degrees of six become technically possible if a molecule happens to have two bifurcated hydrogen bonds at once. With an acceptance angle of 35°, we found about 7% had five bonds and 1% had six bonds and the average degree was 3.3-3.5 at 300 K. Lowering the acceptance angle to 30° removed the molecules with degree six, but also lowered the average degree to 2.9, which is lower than the number inferred from experiment (experimentally, the average degree is believed to be ≈ 3.5). However, at 330K, the average degree jumped up to 3.7 unexpectedly with 30°, suggesting something is spurious about our 300 K data. Thus we choose to use the more conservative value of 30° for this work. A detailed analysis of different hydrogen-bond criteria including plots of bond lifetimes and angle distributions is given by Kumar et. al.[9] The code used for hydrogen bond identification and generation of the adjacency matrix was written by my advisor, Prof. Marivi. Fernández-Serra.

4.2 Storing the bond data

One of the decisions one always has to make before doing any network analysis is the format for storing the network. The three most popular formats are the adjacency matrix, adjacency list and adjacency tree, and each format has many pros and cons.[14] In practice, for large networks ($N > 1000$) the matrix format becomes more cumbersome then the adjacency list & adjacency tree, since the size of the matrix goes as N^2 whereas the size of the list and tree structures goes as $2e + e$ where e is the number of edges (links/bonds). For networks like the internet or social networks, which may contain $100,000+$ nodes, (and yet are sparsely connected) the matrix format is completely out of the question, whereas for small chemical networks it may be a logical choice.

4.3 Finding the degree distribution

The degree of molecule i can easily be calculated by summing along the ith row of the adjacency matrix. As we did when considering percolation theory, let us consider the probability that a given molecule has a hydrogen bond at one of its four binding sites to be p. The probability the bond at a given binding site is broken is $1 - p$. Then the probability of a molecule having j bonds is given by:

$$P(j) = \binom{4}{j} p^j (1-p)^{4-j} = \frac{4!}{j!(4-j)!} p^j (1-p)^{4-j} \tag{12}$$

Table 3: Formulae for the number of cycles of a given type.

order of cycle	equation
3	$\frac{1}{2}\sum_i (A^3)_{ii}$
4	$\frac{1}{2}\sum_i \left[(A^4)_{ii} - ((A^2)_{ii})^2\right]$
5	$\frac{1}{2}\sum_i \left[(A^5)_{ii} - 4(A^3)_{ii}(A^2)_{ii}\right]$

We tried fitting to this function at various temperatures. (see below) This binomial distribution works remarkably well even though it ignores the well established phenomena of hydrogen bond cooperativity.

4.4 Finding non-short-circuited polygons

Enumerating all NSCPs of size n is a very computationally intensive task. Rahmann & Stillinger do not provide much detail on how they searched for NSCPs, other than the fact that they stored their matrix in a list, and that finding NSCPs of size greater than 11 was too computationally intensive for them. One must be particularly careful when dealing with molecules at the edge of the simulation cell, since almost all computer simulations use periodic boundary conditions. If a non-closed polygon stretches across the edge of the unit cell into the period image of the cell, it is quite possible that whatever algorithm is being used will mistakenly count extra NSCPS going across the boundary. To correct for this, Rahmann & Stillinger did their analysis by first surrounding their unit cell with eight identical copies and then considering all the molecules in this "supercell" to be independent. The supercell was then considered to have periodic boundary conditions of its own. In this way, the effects of molecules at the boundary were reduced, since the fraction of molecules at the boundary of the larger cell is smaller.

The field of graph theory offers some help towards attacking this problem. The central object in graph theory is the adjacency matrix A_{ij}. In an unweighted undirected graph, $A_{ij} = 1$ if there is a bond/link/edge between nodes i and j and equals zero otherwise. It is well known that $(A^r)_{ij}$ gives the number of *walks* of length r between i and j. A *walk* of length r is defined as a sequence of edges $\{(n_1, n_2), (n_2, n_3), \cdots, (n_{r-1}, r_r)\}$ where $n_1 = i$ and $n_r = j$. This should not be confused with a *path*, which is a sequence of edges from i to j such that each edge is unique.[1] We will also define a *cycle* as a path from i to i. The term *closed walk* refers to a walk from i to i.

Looking at A_{ii} will give some idea of the number of polygons of size r , but it will also result in a large number of superfluous walks being counted. Let us consider several small values of r. $(A^2)_{ii}$ will give us the number of walks of length two from i to i, which is equal the degree of i. $(A^3)_{ii}$ will give the number of triangles connected to node i - but double counted because both clockwise and counterclockwise are counted. With $(A^4)_{ii}$ things start to get more complicated. $(A^4)_{ii}$ will contain contributions from squares (double counted), but also from walks around the nearest neighbors. It is possible to determine how many such walks are possible – for a node with degree k, k^2 such walks are possible. Now let's consider $(A^5)_{ii}$ - this will include all the pentagons (double counted) but also $2\times$(no of triangles)$\times$(degree) other walks. Table 3 shows how these results can be used to count the number of triangles, squares and pentagons.

Further formulae could be developed, of course, but they will become more and more complicated. However these formulas do not allow us to isolate the non-short-circuited polygons which we are interested in. In graph theory, the proper term for a NSCP is a "chordless cycle" (also called a "graph hole"). A chord of a cycle C is defined as an edge not lying in the edge set of C whose endpoints lie in the vertex set. Counting cycles is well known to be an NP-hard problem, so counting chordless cycles is at least NP-hard and likely belongs to an even more difficult class of problems called #P-hard.

[1]Note: this is a major source of confusion because a quick glance at various websites will show that the term "path" is used differently by different authors. The MIT networks course on Open Courseware advocates using the term "path" for a walk with no repeating edges, but many other references (particularly in older literature) use the term "path" to mean a walk, and the term "simple path" to refer to the case where there are no repeated edges. It appears that the terms "path" and "simple path" were originally in use, but in a quest to make scientific/mathematical literature as dense and abstruse as possible, many authors started not including the qualifier "simple" and left it implied. The term "walk" was introduced in the 1970s to help remedy all the confusion and should be employed in the opinion of this author.

Table 4: Network diameters & cluster sizes. Data from two snapshots at 300K are shown.

Temperature	Network Diameter	Component Sizes
220	11	512
240	12	512
270	11	512
300	12	512
300	18	506, 1
330	12	512
370	13	512
400	12	512

5 Network properties of the H-bond network

5.1 Visualization

The network can be visualized by importing the adjacency matrix into a software package such as *Mathematica*. Here we have plotted the network at 300 K using the 'spring-electrical embedding' feature in *Mathematica*. Spring electrical embedding considers all nodes to be connected by springs and to possess negative electrical charge. The spring and electrical energy function is minimized to make the graph look pretty! (And also to assist in the identification of outlying nodes and various structures in the graph.)

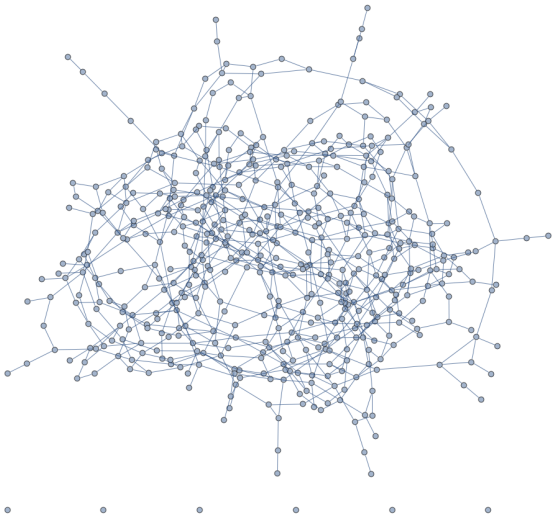

Figure 6: Spring electrical embedding at 300 K. All of the other snapshots we looked at were completely connected. This one at 300 K was interesting because there are some disconnected nodes.

5.2 Network Diameter & Component Sizes

The network diameter is defined as the longest minimal path (geodesic) between two nodes on the network. Technically for disconnected graphs the network diameter is infinite, but it can be redefined as the longest minimal path among the various components of the graph. The size of a component is simply the number of nodes in the component. We found almost all the networks were fully connected, and remained completely connected even after changing the maximum angle in our H-bond definition from 35° to 30°.

5.3 Degree distribution

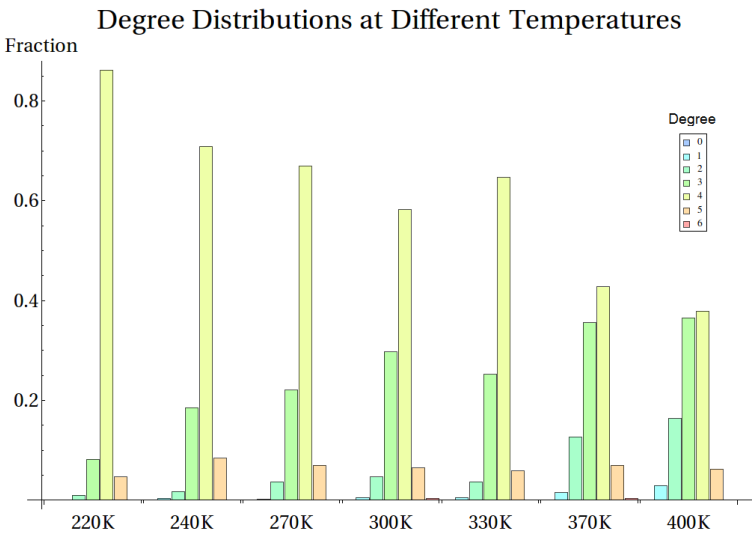

Figure 7: Degree distributions at different temperatures

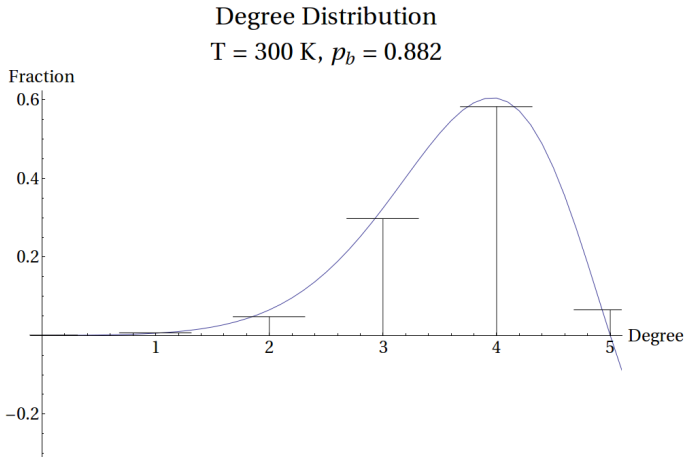

Figure 8: Degree distribution at 300 K fit to a binomial distribution (eqn. 12). The fit was only done between 0 and 4, but the degree of 5 was plotted anyway.

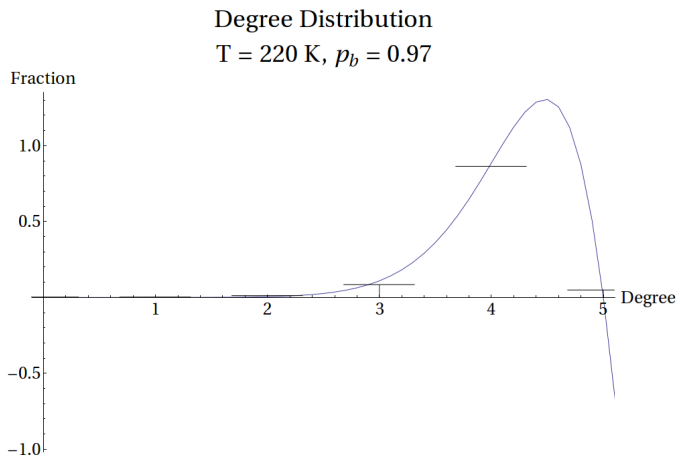

Figure 9: Degree distribution at 220 K fit to a binomial distribution (eqn. 12). The fit was only done between 0 and 4, but the degree of 5 was plotted anyway.

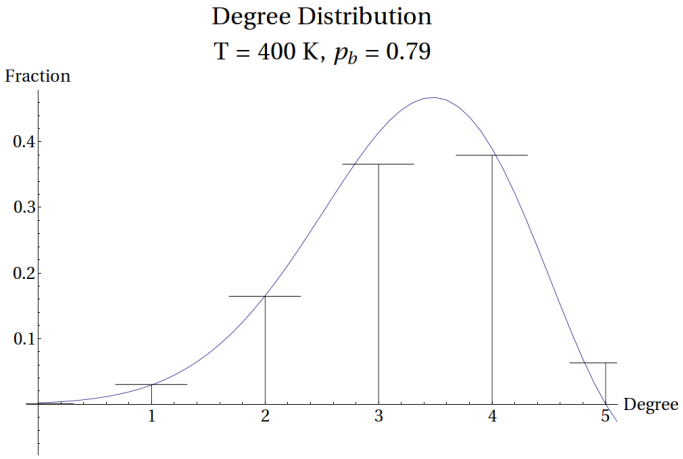

Figure 10: Degree distribution at 400 K fit to a binomial distribution (eqn. 12). The fit was only done between 0 and 4, but the degree of 5 was plotted anyway.

5.4 Cycles

As was mentioned, writing a code to find NSCPs (chordless cycles) is a somewhat painstaking task. However, the *Combinatorica* package in *Mathematica* contains a function called ExtractCycles[] which creates a maximally sized list of disjoint cycles on a graph. A list of disjoint cycles is not at all a complete list of all the cycles, but rather is a list of cycles where no two cycles share an edge. For example at 300 K this function found 47 cycles. This is a small number, but the distribution of sizes gives us some idea of the actual distribution and looks rather similar to distributions which have been previously published:

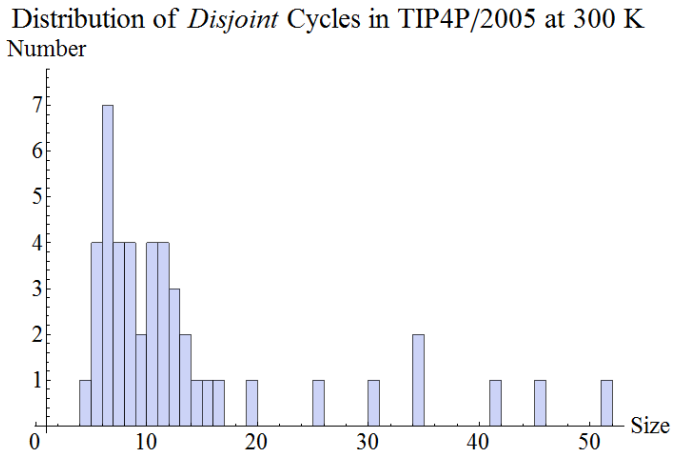

Figure 11: Distribution of the sizes of disjoint cycles at 300 K.

5.5 Clustering coefficients

Then the clustering coefficient of node i is defined as:

$$CC_i = \frac{\text{of pairs of neighbors of } i \text{ which are connected}}{\text{of pairs of neighbors of } i} \tag{13}$$

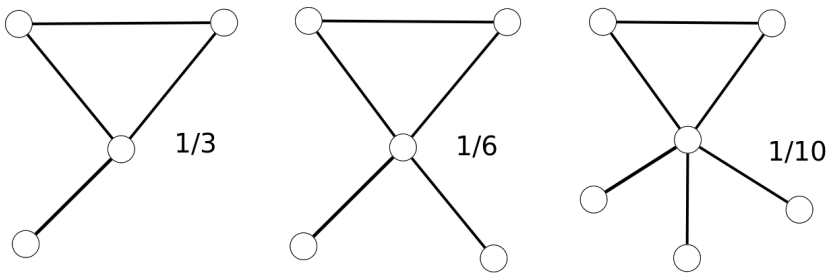

Figure 12: Relevant examples of clustering coefficients for the central node.

One can also define a *average* clustering coefficient as

$$\overline{CC} = \frac{1}{n}\sum_i CC_i \tag{14}$$

One can also define a *global clustering coefficient* as

$$CC = \frac{3(\text{number of triangles})}{\text{number of triplets}} \tag{15}$$

Note that $CC \neq \overline{CC}$ which can be a source of confusion.

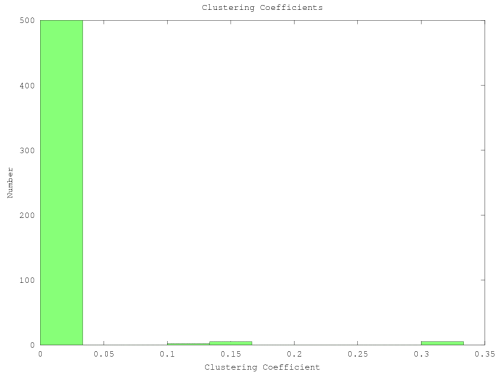

Figure 13: Clustering coefficients at $T = 300$ K. Most molecules had CCs of zero, but there also a few with values of 1/6, 1/3 and 1/10.

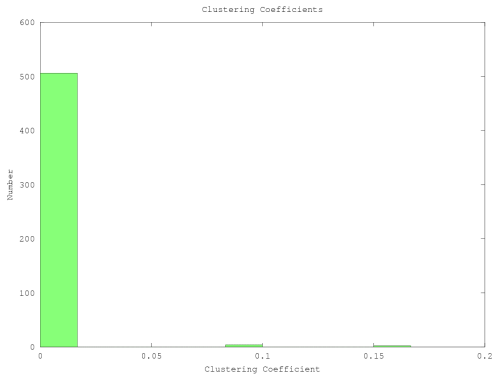

Figure 14: Clustering coefficients at $T = 220$ K. Most molecules had CCs of zero, but there also a few with values of 1/6 and 1/10.

For water we find that most nodes have a clustering coefficient of zero and that clustering is quite low. This is expected because the triangular arrangement is difficult to achieve geometrically.

5.6 PageRank (TM)

Recently it was shown that the PageRank centrality measure can be useful for classifying polyhedral arrangements of molecules, particularly in the liquid phase, where many such arrangements may be possible.[8][11] For instance, water molecules may form polyhedra around dissolved molecules in complex way. For each polyhedra, there is a unique PageRank value, so the PageRank measure provides an easy way of classifying what polyhedra are present in the system given the hydrogen bond adjacency matrix (or some other type of 'bonding' matrix).

PageRank is a measure of centrality, which means it measures how important a given node is. It was originally developed in the context of HTML networks, which are directed networks (so each node has both "in" and "out" degrees), yet the PageRank formula can be applied to undirected networks as well. The equation for the PageRank x_i of node i is:

$$x_i = \alpha \sum_i A_{ij} \frac{x_j}{\text{Max}(k_j^{\text{out}}, 1)} + \beta \tag{16}$$

(The max function must be included to prevent division by zero in the case that a node has no out degrees). β is an intrinsic PageRank which is initially given to each node. If $\beta = 0$ then we will have the issue that a node with zero out-degree will have zero initial PageRank, and thus not contribute to the PageRank of other nodes, which doesn't make much sense. Often times β is set to one. α is an adjustable parameter which is pretty much arbitrary but which should be less than $1/k$, where k is the largest eigenvalue of A_{ij}, to avoid PageRank singularities.[14] In the case

of web search α is interpreted as the probability that a user will click on a link, and β is set as $\beta = (1-\alpha)/N$. The decision to include a $1/N$ in the choice of β was one done simply to ensure that all the PageRanks sum to one, thus giving a probability distribution. The PageRank then can be interpreted as follows: supposing a surfer selects a page out of N pages randomly, and then clicks on links with probability α, then the PageRank of page i gives the probability that the surfer will end up on page i after a long time. Google uses $\alpha = .85$ and for consistency this is the value which is usually used. Mooney experimented with values of $\alpha = .85$ and $\alpha = .99$ in their work.

There are two ways to solve equation 16: iteratively and by a direct matrix solution. The iterative method is usually employed in real-world applications because it is much faster, but in our case we choose to use the direct matrix solution for simplicity. The matrix solution of 16 is:

$$\mathbf{x} = \beta(\mathbf{I} - \alpha\mathbf{A}\mathbf{D})^{-1}\mathbf{1} \tag{17}$$

Here $\mathbf{I}$ is the identity matrix , $\mathbf{1}$ is a column vector of ones and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \frac{1}{\mathrm{Max}(k_j^{\mathrm{out}},1)}$

Number of vertices	Polygon name	Shape	PageRank
4	Square		0.2441558
4	Tetrahedron		0.200000
5	Square pyramid		0.1892430
5	Trigonal bipyramid		0.1772388
5	Wedge		0.2035064
6	Octagon		0.1636142
6	Pentagonal pyramid		0.1822820
6	Trigonal prism		0.1929308

J Math Chem (2012) 50:2342–2350

Figure 15: PageRanks for the central molecule (solute molecule) for different polyhedral arrangements. By computing the PageRank for each solute particle and the molecules within a radius r of the particle, PageRank can be used to efficiently classify the types of polyhedra which surround the solute particles. Figure taken from [8].

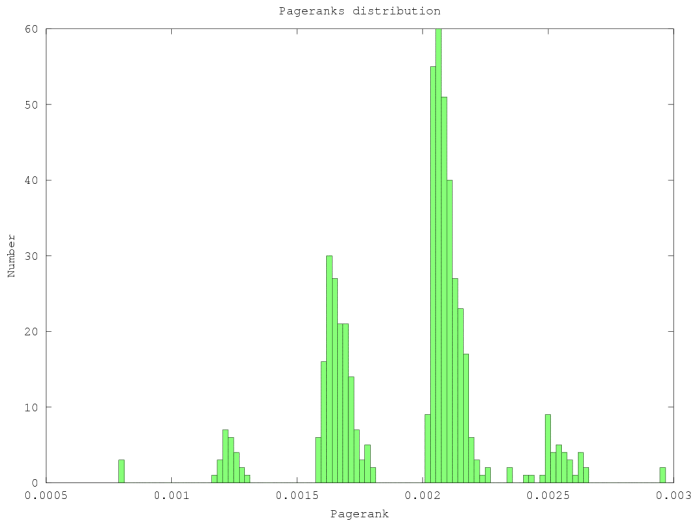

Figure 16: PageRanks at $T = 300$ K with $\alpha = .85$, $\beta = (1 - \alpha)/N$. As is fairly obvious, the peaks correspond to sets of waters with $\approx$ 0, 1, 2, 3, 4, and 5 hydrogen bonds. The number of hydrogen bonds of neighboring waters (and overall structure of the network) is also important, thus explaining why the peaks have some breadth.

6 Possible future work

There is some future work which could be done:

- Write code to find the number of NSCPs / chordless cycles.
- Check out other hydrogen bond criteria besides distance/angle.
- See if averaging over a few frames or larger amount of data helps make the graphs nicer. (averaging over frames is standard practice. I only used one snapshot.)

References

[1] Desiraju G.R. et. al Arunan, E. Definition of the hydrogen bond (iupac recommendations 2011). *Pure Appl. Chem.*, 83:1637, 2011.

[2] Alan C. Belch and Stuart A. Rice. The distribution of rings of hydrogen-bonded molecules in a model of liquid water. *The Journal of Chemical Physics*, 86(10):5676–5682, 1987.

[3] Robin L. Blumberg, H. Eugene Stanley, Alfons Geiger, and Peter Mausbach. Connectivity of hydrogen bonds in liquid water. *The Journal of Chemical Physics*, 80(10):5230–5241, 1984.

[4] M. Chaplin. Anomalous properties of water, Nov 2012.

[5] V. N. Chukanov. Is percolation of relevance to the superheating of light and heavy water? *The Journal of Chemical Physics*, 83(4):1902–1908, 1985.

[6] A. Geiger, F. H. Stillinger, and A. Rahman. Aspects of the percolation process for hydrogen-bond networks in water. *The Journal of Chemical Physics*, 70(9):4185–4193, 1979.

[7] T Head-Gordon. Is water structure around hydrophobic groups clathrate-like? *Proceedings of the National Academy of Sciences*, 92(18):8308–8312, 1995.

[8] Matthew Hudelson, Barbara Logan Mooney, and Aurora E. Clark. Determining polyhedral arrangements of atoms using pagerank. *Journal of Mathematical Chemistry*, 50:2342–2350, 2012.

[9] R. Kumar, J. R. Schmidt, and J. L. Skinner. Hydrogen bonding definitions and dynamics in liquid water. *The Journal of Chemical Physics*, 126(20):204107, 2007.

[10] Yang Liu, Athanassios Z. Panagiotopoulos, and Pablo G. Debenedetti. Low-temperature fluid-phase behavior of st2 water. *The Journal of Chemical Physics*, 131(10):104508, 2009.

[11] Barbara Logan Mooney, L.Ren Corrales, and Aurora E. Clark. Molecularnetworks: An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation. *Journal of Computational Chemistry*, 33(8):853–860, 2012.

[12] Kazi A. Motakabbir and M. Berkowitz. Isothermal compressibility of spc/e water. *The Journal of Physical Chemistry*, 94(21):8359–8362, 1990.

[13] O.A. Nabokov and Yu.A. Lubimov. The dielectric relaxation and the percolation model of water. *Molecular Physics*, 65(6):1473–1482, 1988.

[14] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[15] Lívia Pártay and Pál Jedlovszky. Line of percolation in supercritical water. *The Journal of Chemical Physics*, 123(2):024502, 2005.

[16] A. Rahman and F. H. Stillinger. Hydrogen-bond patterns in liquid water. *Journal of the American Chemical Society*, 95(24):7943–7948, 1973.

[17] Stuart A. Rice and Mark G. Sceats. A random network model for water. *The Journal of Physical Chemistry*, 85(9):1108–1119, 1981.

[18] Mark G. Sceats and Stuart A. Rice. The enthalpy and heat capacity of liquid water and the ice polymorphs from a random network model. *The Journal of Chemical Physics*, 72(5):3248–3259, 1980.

[19] Mark G. Sceats and Stuart A. Rice. The entropy of liquid water from the random network model. *The Journal of Chemical Physics*, 72(5):3260–3262, 1980.

[20] Mark G. Sceats and Stuart A. Rice. A random network model calculation of the free energy of liquid water. *The Journal of Chemical Physics*, 72(11):6183–6191, 1980.

[21] Mark G. Sceats, M. Stavola, and Stuart A. Rice. A zeroth order random network model of liquid water. *The Journal of Chemical Physics*, 70(8):3927–3938, 1979.

[22] Eli Shiratani and Masaki Sasai. Growth and collapse of structural patterns in the hydrogen bond network in liquid water. *The Journal of Chemical Physics*, 104(19):7671–7680, 1996.

[23] Robin J. Speedy. Self-replicating structures in water. *The Journal of Physical Chemistry*, 88(15):3364–3373, 1984.

[24] Robin J. Speedy and Mihaly Mezei. Pentagon-pentagon correlations in water. *The Journal of Physical Chemistry*, 89(1):171–175, 1985.

[25] H. Eugene Stanley and J. Teixeira. Interpretation of the unusual behavior of h2o and d2o at low temperatures: Tests of a percolation model. *The Journal of Chemical Physics*, 73(7):3404–3422, 1980.

[26] Frank H. Stillinger and Aneesur Rahman. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, 60(4):1545–1557, 1974.

[27] Frank H. Stillinger and Aneesur Rahman. Molecular dynamics study of liquid water under high compression. *The Journal of Chemical Physics*, 61(12):4973–4980, 1974.

[28] Walter H. Stockmayer. Theory of molecular size distribution and gel formation in branched-chain polymers. *The Journal of Chemical Physics*, 11(2):45–55, 1943.