Cooking Beyond Your Front Door: Novel Fusion Recipe Generation

Tiarnán Finn-James Curran-Feeney[a,1]

[a]Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL

This manuscript was compiled on May 1, 2022

Here we describe a novel recipe generation using an ingredient network generated from the CulinaryDB database. Focusing on fusion recipes, we will study the underlying block structure and use a modified stochastic block model to construct novel fusion recipes.

Recipe Generation | Recipe Network | Block Structure | Random Walk

The introduction of the world wide web has allowed for rapid growth in the total number of recipes an individual has access to. Through this rapid access to data the popularity of fusion recipes, where a recipe takes ideas from two or more traditional cuisines, has grown. Over the past decade, the freely available data has been analysed with networks to attempt to discover the underlying connections between ingredients in recipes.

Two prime example of this work are (1) and (2), both of which focus on ingredient pairings. They investigated the popular food pairing hypothesis, whereby two similar tasting are more likely to be used together in a recipe which is known as positive food pairing. Results from (1) found that Jordan, Lebanon, Syria, and Palestine recipes supported the food pairing hypothesis. However, (2) found that while the strength of the negative pairing varied by region in India, all regions within India have recipes that demonstrate strong negative food pairings in contrast to the food pairing hypothesis. Other examples of such work include the focus on the dynamics of the co-occurrence of different ingredients during different seasons of the year (3).

Work has also been done to generate recipe recommendation systems. One such example is (4), which used a network to generate ingredient recommendations based on an initial input ingredient. This paper contributes to the field by looking at novel recipe generation with respect to regional and fusion cuisines. Through uncovering the underlying block structure of an ingredient network, we intend to generate a new network that can then be explored using a modified random walk to generate novel recipes. Through novel fusion recipes, we intend to bring network science into the modern cooking world.

Results

We analysesd a 6 layer network in which, as described in **Materials and Methods: Network Construction**, each layer represents a region. We have included the regions Africa, Britain and Ireland, Eastern Europe, the Indian Subcontinent, Korea, and South America.

By performing the weighted stochastic block model, which is described in **Materials and Methods: Network Block Structure**, on our network, with $\beta = 1$ and $\alpha = 0.5$, we obtained a block structure for the network. Figure 1 illustrates the change in the adjacency matrix for South America by grouping nodes into their blocks. In the right hand image, we can see that there is a core selection of ingredients that are frequently used together, forming a core periphery. Outside of that core, ingredient pairings are less common. There are some weak connections between ingredients outside the core to ingredients within the core. We also observe a second core with weaker connections between itself and stronger connections to the original core.

Across all 6 layers we see a similar behaviour. The number of nodes comprising the two cores, as well as the strength of the connections within and between the cores varies for each layer. For example, Figure 5 illustrates the block structure for Korea's layer where one of the cores is comprised of only three nodes. Additionally, while there were 1,033 possible ingredients recorded in the database, each region uses less than 400 distinct ingredients in their respective recorded recipes.

In order to construct novel recipes, we will need to choose γ, the probability of teleportation, and L, the length of the walk, so that the distribution of the number of ingredients in a recipe matches the distribution of recipes in CulinaryDB. The distribution skews heavily towards shorter ingredient list lengths when γ is set to zero. This is illustrated in Figure 7, where we can see the distribution for the modified random walk around the layer representing Africa. Through varying L and γ, we settled on 30 and 0.1, respectively. Figures 7-12 illustrate how the distribution of recipe ingredient list lengths under the chosen parameter value compares to the true distribution of recipe ingredient list lengths in CulinaryDB for each region. We chose γ and L so as to minimise the total difference between the two distributions for each of the 6 regions.

Table 1 contains a sampled selection of novel recipes that were generated by the modified random walk for different δ values. The parameter δ is the probability of switching layer before selecting the next ingredient.

Significance Statement

Network Science can be used to examine the structure of a data set that cannot be seen by the naked eye. With the introduction of online recipe blogs, a large repertoire of regional cuisines is available to study and in the past decade there has been an interest in investigating ingredient networks. Through using a modified Stochastic Block Model, we intend to build upon the work previously done in order to generate novel fusion recipes between a selection of different regional cuisines.

Network constructed and investigated by, report written by, and reviewed by Tiarnán Finn-James Curran-Feeney.

No conflict of interest known.

[1]To whom correspondence should be addressed. E-mail: tiarnan.curran-feeney@warwick.ac.uk

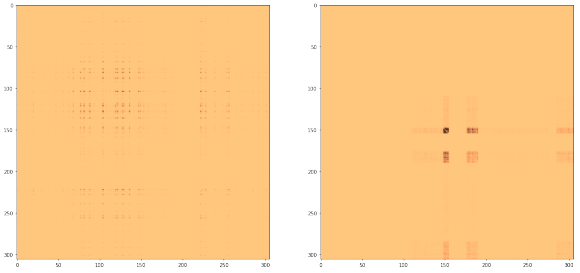

Fig. 1. Plot of the adjacency matrix for the layer representing Africa. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any African recipes have been excluded

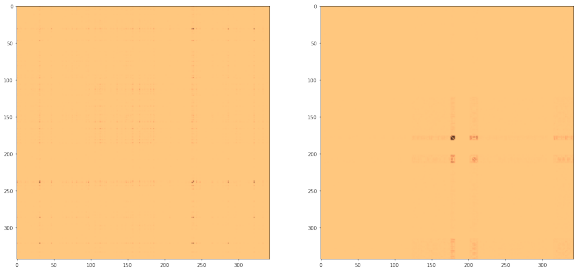

Fig. 2. Plot of the adjacency matrix for the layer representing Britain and Ireland. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any recipes in Britain and Ireland have been excluded

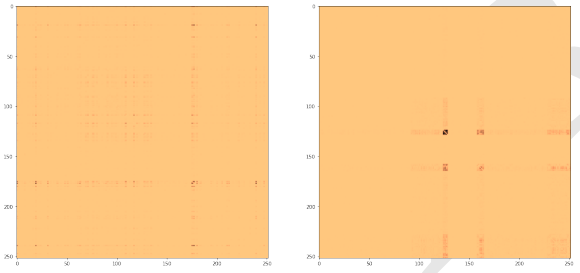

Fig. 3. Plot of the adjacency matrix for the layer representing Eastern Europe. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any Eastern European recipes have been excluded.

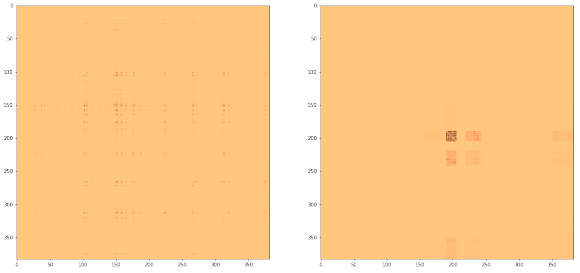

Fig. 4. Plot of the adjacency matrix for the layer representing Indian Subcontinent. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any recipes from the Indian Subcontinent have been excluded.

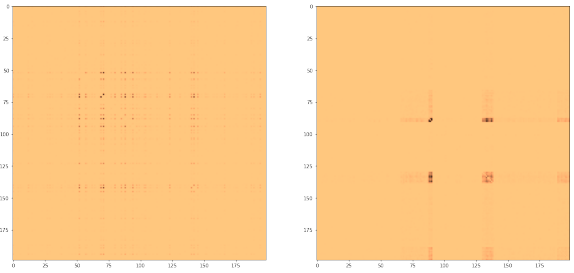

Fig. 5. Plot of the adjacency matrix for the layer representing Korea. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any Korean recipes have been excluded.

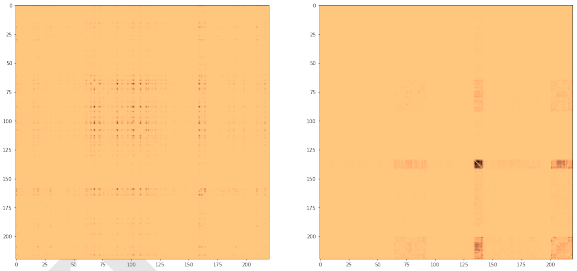

Fig. 6. Plot of the adjacency matrix for the layer representing South America. Left is the original adjacency matrix and right is the adjacency matrix when nodes have been reordered into the groups found by the WSBM. Nodes that represent ingredients not used in any South American recipes have been excluded.

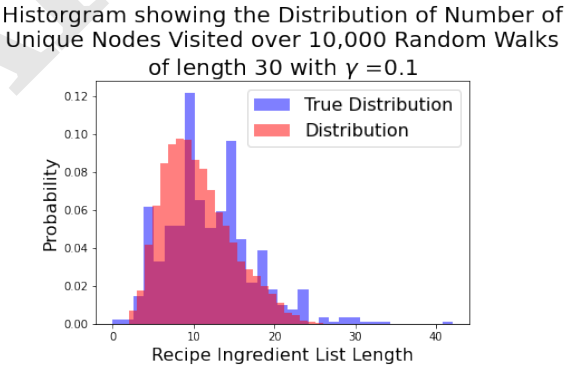

Fig. 7. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for Africa.

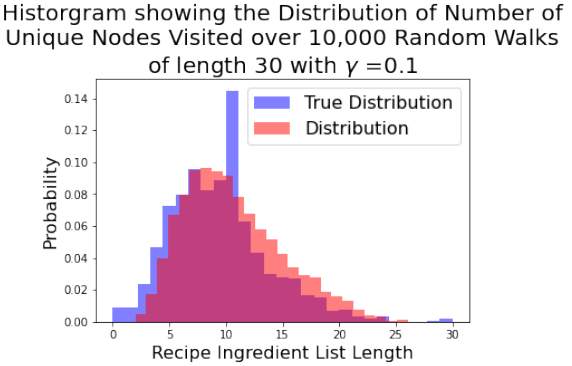

Fig. 8. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for Britain and Ireland.

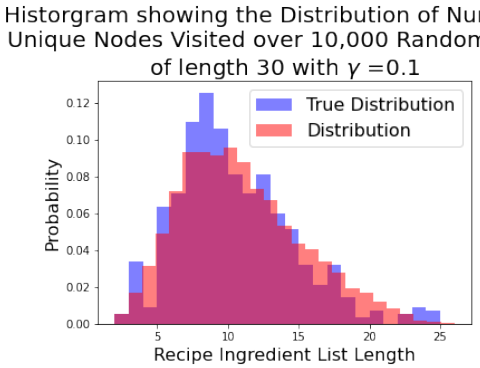

Fig. 9. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for Eastern Europe.

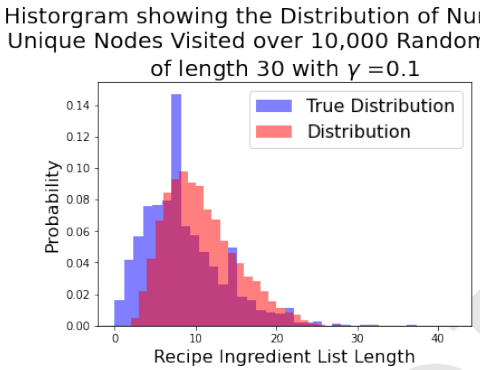

Fig. 10. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for the Indian Subcontinent.

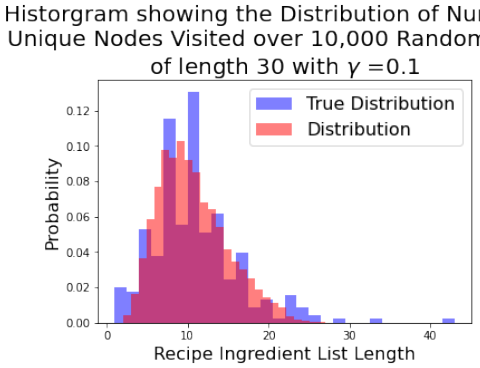

Fig. 11. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for Korea.

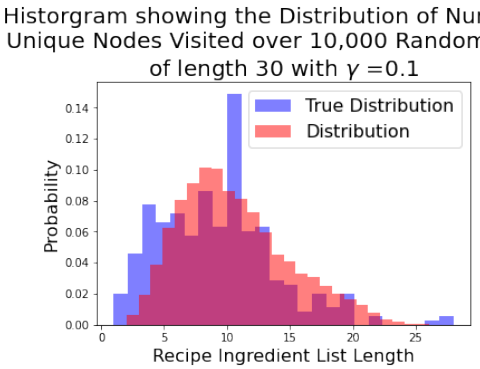

Fig. 12. Histogram plotting the distribution of recipe ingredient length for recipes generated by our modified random walk (in red) and the distribution in CulinaryDB (in blue) for South America.

Table 1. A sample of novel recipes generated by the modified random walk.

δ	Starting Region	Ingredient List		
0	Britain/Ireland	Chickpea	Almond	Coconut Milk
		White Currant	Sassafras	Basil
		Cooking Oil		
0.1	Africa	Bread	Sweet Potato	Aubergine
		Couscous	Almond	Tomato
		Rice	Yogurt	Cream
		Olive	Mint	Cayenne
		Laurel	Cinnamon	Cumin
		Honey	Ginger	Oregano
		Salt	Water	
0.5	South America	Meatball	Shiitake	Mutton
		Hazelnut	Oregano	Fettucine
		Lotus	Persimmon	Coriander seed
		Jujube	Couscous	Rice
		Peanut Butter	Butter	Cinnamon
		Barbeque Sauce	Goat Milk	Thistle
		Corn Grits	Sesbania Flower	Garlic

Discussion

Through a modified random walk, we have produced a way to generate novel recipes from a given region whose ingredient list lengths follow a similar distribution to that of a recipe from that region in the database CulinaryDB. Additionally, our method allows for the implementation of novel fusion recipes that incorporate characteristics from a selection of different regions.

Of the 1,033 ingredients in the CulinaryDB database, none of the 6 regions we considered used more than 400 distinct ingredients. The number of ingredients used varied depending on the region. We have seen that all regions have a core selection of ingredients that form the basis of their cuisine that other ingredients may be paired with. Additionally, regions have a secondary core that pairs strongly with the primary core and has slightly weaker connections within itself. This structure mirrors cuisines in the real world. If we take baking as an example, we would expect a strong core of flour, sugar, eggs, water, oil, and butter. We would also expect to see a secondary core of flavourings, such as chocolate, vanilla pods, coconut, etc... This secondary core would have strong connections to the primary core but they would less frequently be used with each other.

One of the weaknesses in the WSBM is that real world networks typically have long-tail degree distributions, which is an effect the WSBM does not reproduce. In (5), Aicher et al. investigated the WSBM and proposed that we could implement the work done by Kerrer B. et al. in (6), which uses degree correction to force the long-tail degree distribution structure. However, Aicher C. et al. found that the degree corrected WSBM did not perform as well as the WSBM when recovering the underlying structure of a network. Therefore, using the WSBM likely performs well enough for the task at hand. An intended extension of this project is to investigate the network using the degree corrected WSBM and compare the results found.

Britain and Ireland have vastly different population sizes. The combined populations of the Republic of Ireland and Northern Ireland sum to 6.88M, while London alone comprises a population of 8.982M which is over 2M more than the whole of Ireland. This likely leads to a disproportionate British influence over the British-Irish layer. As a result, the layer may not be very revealing of Irish cuisine. However, due to the close geographical nature of the two Islands and the strong historical connections, we have assumed that this has not influenced the results too heavily.

The Korean layer focuses on a single culture and likely results in an accurate representation of the regional cuisine. However, the other layers (excluding Britain and Ireland) cover large portions of the world where several cultures are represented in a single layer. In particular, Africa is comprised of 54 countries and this doesn't even touch upon the number of distinct cultures within these countries. The distinction between the cultural cuisines is lost by the classification of the entire continent as one region. As a result, the layer may not replicate the true structure of African cuisine.

It is because of the internet that databases like CulinaryDB were possible to compile. However, the database is comprised of modern recipes from cooking websites. With the interconnected world that the internet has created, modern recipes from a given cuisine can have heavy influence from various regions accross the world. As such, the accuracy to which the recipes represent the traditional cuisine is unknown. The work could be improved by including the database FlavorDB, (7), which is designed to be used alongside CulinaryDB. It is a database listing the key flavour compounds an ingredient is made from. By using research on the food pairing hypothesis, such as (1) and (2), we could build upon our fusion model by considering whether a region traditionally follows a positive or negative food pairing in their recipes.

CulinaryDB offers recipes from 22 regions, excluding miscellaneous regions with only a small number of recipes, while we only availed of 6 regions. The work in this paper can easily be expanded to incorporate all 22 regions. This would allow for a more complete network that is capable of a greater number of fusion options in novel recipe generation.

Our constructed network did not distinguish between recipe types, such as sweet, savoury, drink, or sour. This resulted in a difficulty to produce feasible sweet recipes as key dessert ingredients, such as egg, flour, and sugar, are also regularly used in savoury food. The unintended consequence is an unwanted combination of sweet and savoury. An intended expansion to this project is to analyse a similar network with layers that distinguish between recipe type rather than region.

Our modified random walk constructs a proposal for the ingredient list of a novel fusion recipe. A final intended expansion in future research will be to attempt to expand upon this method to include potential cooking techniques for the novel fusion recipes. A proposal for how this might be done is to construct a bipartite network of ingredients and cooking techniques with connections between nodes based on how frequently an ingredient and a cooking technique occur in the same recipe. This would allow for an interesting study into the connections between food and prepatory methods.

Materials and Methods

CulinaryDB is a database comprised of 45,722 recipes denoting the ingredients contained in each one and the region of the cuisine. The recipes were collected from a range of sites, including Tarla Dalal, AllRecipes, and Epicurious. By constructing a weighted network, we will attempt to generate novel recipes from a selection of regional cuisines. CulinaryDB is a publically available database and can be found here. All code produced will be available in Supplementary Information section 2.

Network Construction. We will begin by considering each region separately and construct separate networks for each. By representing each ingredient as a node, we wish to incorporate the relative importance of that ingredient in the regional cuisine and the probability of two ingredients being used together. To do this, each edge (a,b) for region R will be weighted with some weight $W_R(a,b)$. In (8), Teng CY et al. construct an undirect weighted ingredient network using the weights shown in 1, where $p_R(a_1, a_2, \ldots, a_n)$ is the probability that given a recipe at random from region R it will contain the ingredients $a_1, \ldots, a_n$. While Teng CY et al. did not consider layered networks, we have included the subscript R for consistent notation within this report.

$$W_R(a,b) = log(\frac{p_R(a,b)}{p_R(a)p_R(b)}) \quad [1]$$

The key drawback of this method is that it is undirected. In real life, the likelihood of ingredient b being in a recipe given that a is in the recipe is not necessarily equivalent to the likelihood of ingredient a being in a recipe given ingredient b is present. For example, take flour and yeast. The vast majority of recipes involving yeast will

also include flour. However, a recipe involving flour could contain baking powder, bicarbonate of soda, yeast, or no rising agent at all. We can account for this by modifying our network to a directed network, see Equation 2, where $N_R(a_1, \ldots, a_n)$ is the number of recipes in CulinaryDB for region R which contain the ingredients $a_1, \ldots, a_n$. The weight of an edge, $W_R(a, b)$, now represents the logarithm of the probability that b is in a recipe given that a is.

$$W_R(a,b) = log(\frac{p_R(a,b)}{p_R(a)}) \\ = log(N_R(a,b)) - log(N_R(a)) \quad [2]$$

While an improvement, this new ingredient network does not yet express the relative importance of an ingredient in the cuisine. We have no way of working out the frequency with which an ingredient is used in a regions's cuisine from our network. We can account for this by reducing back down to an undirected network with the new weights in Equation 3. The frequency with which two ingredients occur together now represents the edge weight. We cannot recover the probability of ingredient b given ingredient a is in the recipe but we can calculate the relative importance of ingredient b relative to ingredient a as $W_R(a,b)/(\sum_c W_R(a,c))$. Meanwhile the relative importance of an ingredient can be represented by the sum of edge weights that come from it.

$$W_R(a,b) = N(a,b) \quad [3]$$

In order to connect our network layers together we will introduce a parameter β. Each node in a layer will be connected to its respective counterpart in another layer with weight β. I.e. all nodes that represent the same ingredient will form a cluster. By varying the strength of β, we will be able to explore both regional block structure and global block structure. We will be taking $\beta = 1$ in this paper. A visualisation of a simple example network is shown in Figure 13.

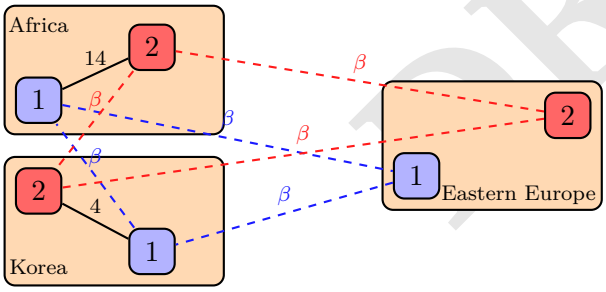

Fig. 13. An illustration of an ingredient network with two ingredients and three layers.

Network Block Structure. In order to generate novel recipes we are interested in the underlying structure of our network. We will be using a stochastic block model (SBM) to do this. An SBM will categorise each node into one of k possible groups, were k is a known constant. The normal SBM is designed for unweighted networks. We could adapt our network to be suitable for SBM by considering only the edges with a weight greater than some threshold value. However, in doing so we lose a lot of the important information that our network tells us about the co-occurrence of ingredients as well as the popularity of an ingredient in a regional cuisine.

In (9), Aicher C. et al. first introduced their modified SBM to approach weighted networks called the Weighted SBM (WSBM). To explain how it works we will first consider the regular SBM. The SBM returns a $k \times k$ matrix, θ, of parameter values whereby θ_{ij} is the probability of an edge existing between a node in group i and a node in group j and a vector z with z_i equal to the group assignment of node i. θ and z are found by the SBM maximising the likelihood function, Equation 4, where A is the $n \times n$ adjacency matrix for the network with entries $A_{ij} \in \{0, 1\}$.

$$Pr_E(A|z,\theta) = \prod_{ij} \theta_{z_i z_j}^{A_{ij}} (1 - \theta_{z_i z_j})^{1 - A_{ij}} \quad [4]$$

In order to account for weighted networks in the WSBM, we introduce a tuning parameter α. If we denote the likelihood function for the edge weights as $Pr_W(A|z,\theta)$, we can create a new likelihood function considering both the edge existence and edge weight, Equation 5.

$$log(Pr(A|z,\theta)) = \alpha log(Pr_E(\hat{A}|z,\theta_E)) \\ + (1-\alpha) log(Pr_W(A|z,\theta_W)) \quad [5] \\ \hat{A}_{ij} = \mathbb{1}_{\{A_{ij}>0\}}$$

We assume that if we subtract one from each of our edge weights they will follow a poisson distribution. In doing so we get the likelihood function, Equation 6. We will be following the recommendation of Aicher et al, in (5), and we will take $\alpha = 0.5$.

$$log(Pr_W(A|z,\theta_W)) = \sum_{i,j,A_{ij}>0} (A_{ij} - 1) log(\theta_{W z_i, z_j}) - \theta_{W z_i, z_j} \\ - \sum_{k \in \{1,2,\ldots,A_{ij}-1\}} log(k) \quad [6]$$

The WSBM makes the assumption that we know the exact number of groups, K. We will be approximating K by using a strategy implemented in (10). We will choose a random layer of our graph and run the WSBM on it for a selection of different K values. The smaller testing graph will allow the WSBM to run faster. We will then choose the value of K that maximised the likelihood function for the testing graph. As Figure 14 shows, the increase in the log-likelihood plateaus as we increase K. Taking $K = 10$ should be sufficient.

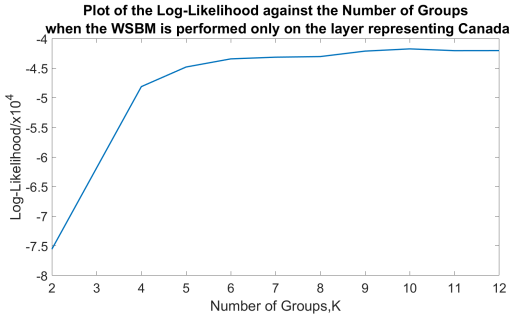

Fig. 14. Plot of the Log-Likelihood against the Number of Groups, K for the network representing the region Canada.

Recipe Generation Strategy. Using the underlying block structure discovered, we will construct a new ingredient network. This will allow for the possibility to discover new connections between ingredients that did not appear in the original network. We will explore our newly constructed network using a random walk with teleportation. The traditional random walk starts at a random node and will travel to a new node along an edge with some known probability based on the current node. The addition of teleportation means that with some probability ν, rather than travelling to a nearby node along an edge, the next node will be randomly chosen.

There are two key factors to a recipe that we want to be able to vary, which are the predictability of the ingredients and the level of dependence on the chosen region. the former can be incorporated through teleportation while the latter can be established through a

fixed probability for the edges connecting corresponding ingredients in different regions. We shall introduce two parameters γ and δ, which represent the probability of randomly jumping to a new node and the probability of changing region, respectively.

For the modified random walk, we will say that the probability of travelling from node a to b, both of which represent ingredients in the same region and given no teleportation is proportional, to the frequency with which they occur together, Equation 7. This will allow us to travel around our newly constructed network based on the relative importance of ingredients to the current ingredient.

$$P(x_{n+1} = b|x_n = a, R, \gamma = \delta = 0) = \frac{W_R(x_n, a)}{\sum_b W_R(x_n, a)} \quad [7]$$

We shall denote N, M as the number of ingredients and the number of regions, respectively. Then by introducing teleportation, our probability of travelling to node b is now as shown in Equation 8. However, even without teleportation, this modified random walk may create unrealistic recipes, as if it reaches a popular node it could travel to any part of the network in a short number of steps. For example, if we start at vanilla pods, the next node could be flour because both may be used in a dessert. Flour is also used in savoury foods, through a beef pie we could easily end up at beef. However, vanilla's stubborn refusal to ever appear in a savoury dish throughout history only shows how this would likely make a terrible combination.

$$P(x_{n+1} = b|x_n, R) = (1 - \gamma - \delta)\frac{W_R(x_n, b)}{\sum_a W_R(x_n, a)} + \frac{\gamma}{N} + \frac{\delta}{M-1}\sum_{R' \neq R}\frac{W_{R'}(x_n, b)}{\sum_a W_{R'}(x_n, a)} \quad [8]$$

We shall make one more modification to our random walk with teleportation to overcome this issue of unrealistic ingredients. Rather than taking the next step of our random walk from the most recent node arrived at on our walk, we shall randomly select a node from all nodes that have been visited. We want common ingredients to be less likely to be chosen. To do this, we shall choose the k^{th} node visited, x_k, with probability proportional to the inverse of the sum of all the weights of edges emanating from it, Equation 9. This gives more importance to less frequently used ingredients and if a node is regularly visited then it will be more likely to be selected and will have more importance in the recipe.

$$P(x_k \text{ chosen}) = \frac{1}{F\sum_c W_R(x_k, c)} \quad [9]$$
$$F = \sum_k \frac{1}{\sum_c W_R(x_k, c)}$$

ACKNOWLEDGMENTS. A special thanks to Dr. Marya Bazzi for giving guidance to me on my project outline. Not to mention my housemates Akanksha, Caitlin, Emily, and Liam for being bold enough to try the recipes suggested.

1. Issa L, Alghanim F, Obeid N (2018) Analysis of food pairing in some eastern mediterranean countries in *2018 8th International Conference on Computer Science and Information Technology (CSIT)*. (IEEE), pp. 167–172.
2. Jain A, NK R, Bagler G (2015) Analysis of food pairing in regional cuisines of india. *PloS one* 10(10):e0139539.
3. Kikuchi Y, Kumano M, Kimura M (2017) Analyzing dynamical activities of co-occurrence patterns for cooking ingredients in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. (IEEE), pp. 17–24.
4. Pini A, Hayes J, Upton C, Corcoran M (2019) Ai inspired recipes: Designing computationally creative food combos in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. pp. 1–6.
5. Aicher C, Jacobs AZ, Clauset A (2015) Learning latent block structure in weighted networks. *Journal of Complex Networks* 3(2):221–248.
6. Karrer B, Newman ME (2011) Stochastic blockmodels and community structure in networks. *Physical review E* 83(1):016107.
7. Bagler G, Singh N (2018) Data-driven investigations of culinary patterns in traditional recipes across the world in *2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW)*. (IEEE), pp. 157–162.
8. Teng CY, Lin YR, Adamic LA (2012) Recipe recommendation using ingredient networks in *Proceedings of the 4th annual ACM web science conference*. pp. 298–307.
9. Aicher C, Jacobs AZ, Clauset A (2013) Adapting the stochastic block model to edge-weighted networks. *arXiv preprint arXiv:1305.5782*.
10. Airoldi EM, Blei DM, Fienberg SE, Xing EP, Jaakkola T (2006) Mixed membership stochastic block models for relational data with application to protein-protein interactions in *Proceedings of the international biometrics society annual meeting*. Vol. 15.