Shallow Cumulus Clouds as Complex Networks

by

Pouriya Alinaghi

M.Sc. Thesis

Delft University of Technology

Faculty of Civil Engineering and Geosciences

Department of Geoscience & Remote Sensing

Shallow Cumulus Clouds as Complex Networks

by

Pouriya Alinaghi

to obtain the degree of Master of Science in
Applied Earth Sciences
Track: Environmental Engineering
Specialization: Environmental Science
at the Delft University of Technology,
to be defended publicly on **Wednesday May 26, 2021 at 2:00 PM**.

Student number:	5078946	
Project duration:	November 15, 2020 – May 26 , 2021	
Thesis committee:	Dr. F. Glassmeier,	TU Delft
	Prof. dr. A. P. Siebesma,	TU Delft
	Dr. R. Taormina,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Abstract

Clouds, specifically shallow clouds, are known as a major source of uncertainty in global climate models. Shallow clouds over the global oceans show different spatial patterns and organizations that may be influenced by climate change. Besides, the frequency of these patterns can change the climate feedback of marine clouds in the subtropics. Experts in atmospheric sciences have proposed different kinds of organization metrics. In a recent paper (Janssens et al., 2021), 21 cloud organization metrics were collected and computed over 5000 satellite scenes of shallow clouds located near the east of Barbados. Almost all of the existing metrics of cloud organization are bulk parameters of the cloud field. Accordingly, over the same data as Janssens et al. (2021), this project aims to define a metric that originated from the mutual arrangements of the individual clouds in the field. To this end, network theory as a mathematical tool is employed to define new cloud organization metrics to explore how cloud objects interact with and coordinate concerning each other. It should be noted that this research is the first in which cumulus cloud fields are considered as complex spatial networks.

In this regard, an important question is whether a new network metric can distinctly explain a variability which has not been captured by previously defined metrics in cloud organization. To address the research question, we utilize a multivariate regression model to understand whether a linear combination of principal components of the existing metrics can encapsulate a variation in newly proposed network metrics. We found that degree standard deviation and mean of clustering coefficient are two network metrics that can distinctly capture a variability that has not been encapsulated by the existing metrics in cumulus cloud organizations. Degree standard deviation simultaneously measures the homogeneity of the cloud size distribution and the distance between nearest neighboring clouds distribution. A large value of the average clustering coefficient indicates that the network field consists of either large clouds located on corners of triangles or relatively small clouds closely coagulated. Additionally, two sensitivity analyses were performed to understand how the main results are influenced by either center-spacing or edge-spacing distance between clouds and the nodes (clouds) located close to the boundaries of the field. Finally, one should note that all of the results of network theory-based approaches are comprehensively affected by how the network is defined. Defining a network that combines both geometric and physical interaction between cloud objects will be the future work of this research.

Preface

During my journey as an M.Sc. student in Environmental Engineering, I became familiar with vast new concepts from atmospheric sciences. These concepts have made me enthusiastic about the atmospheric boundary layer and the processes that occur within it. I am happy that I understood physical processes occurring in various scales ranging from the turbulence scale to the global scale. Understanding how Earth is affected by these processes is always motivating for me. On the other hand, since my childhood, I have been a close friend of mathematics, especially the discrete branch of mathematics such as Combinatorics and Graph Theory. Teaching Combinatorics during my B.Sc. can prove my particular interests in this subject. I am lucky that I found an M.Sc. project in which a combination of my interests is included. For the first time, we discovered complex networks on shallow Cumulus cloud fields to investigate cloud organization from another point of view. This study is not the end of my journey, and I will keep on exploring shallow clouds with the networks during my upcoming P.hD. project within the Geoscience & Remote Sensing department of TU Delft.

Huge thanks to Franziska Glassmeier for her patience, accuracy, remarkable availability, tremendous support during every moment of this research ranging from analyses to writing the thesis. Without her loyal help, this thesis could not finish in a comparably short time. I am so grateful to her. To Pier Siebesma for always being ready to supply rigorous feedback, precious advice, and guidance. His overflowing passion and knowledge regarding clouds notably improve this thesis. I am so happy that my P.hD. project will be also supervised by Franziska and Pier. To Riccardo Taormina for being the external member of the committee, dedicating his valuable time to read my thesis, contributing to the meetings, and his responsibility. Also, I would like to thank Martin Janssens for devoting his time to make me familiar with his code, downloading the data from NASA's websites, and his fruitful comments on my work. Besides, I would like to thank all of my teachers in TU Delft particularly Sukanta Basu, Stephan de Roode, Bas van de Wiel, and Louise Nuijens who have provided me invaluable insights into atmospheric sciences.

Last but not least, I am particularly grateful to my family for their endless support during every moment of my life. Especially, with my deepest part of my heart, I appreciate the financial supports of my parents. Being an international student of TU Delft was not possible without their tremendous support. I am proud of my family for dedicating their budget to be spent on the way of science for having a better world, brighter future, and more aware human beings.

Pouriya Alinaghi
Delft, May 2021

Acknowledgment

In this thesis, the cloud masks are downloaded from Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC)(http://dx.doi.org/10.5067/MODIS/MYD06L2.061; http://dx.doi.org/10.5067/MODIS/MOD06L2.061). All downloading and preprocessing of the data is done via the code scripts that were written by Martin Janssens in the paper by Janssens et al. (2021). This code can be found in a GitHub repository (https://github.com/martinjanssens/cloudmetrics). Besides, low-to-high plots of cloud metrics in this research is also performed using the code scripts of Martin Janssens. The network analysis is done using Python and NetworkX package (Hagberg et al., 2008). The processing and post-processing of the data is also done employing Python and its invaluable libraries including Numpy (Harris et al., 2020), Pandas (Wes McKinney, 2010), Scipy (Virtanen et al., 2020), Scikit Learn (Pedregosa et al., 2011), Matplotlib (Hunter, 2007), and Seaborn (Waskom, 2021). Additionally, particular thanks to Franziska Glassmeier for providing the software for how to define a graph on a cloud field (Glassmeier and Feingold, 2017), and for the computation of the edge-spacing distances between cloud objects.

Contents

List of Figures

1 Introduction

Shallow Cumuli are affected by numerous processes. Most of these processes are not resolved in Global Climate Models (GCMs), since the behavior of these processes is extremely difficult to be parameterized (Nuijens and Siebesma, 2019). Also, the cloud radiative effect (CRE) of low clouds plays a crucial role in the global radiation budget of our planet Earth. Although clouds cool the Earth by reflecting the shortwave radiation to space, they can contribute to warm the Earth by trapping the outgoing infrared radiation emitted from the Earth's surface and radiate it backward as a downwelling longwave radiation towards the surface of the Earth (Wallace and Hobbs, 2006). The response of a GCM to the CO_2 doubling is quantified by Equilibrium Climate Sensitivity (ECS) (Nuijens and Siebesma, 2019). These interacting processes have been indicated in the determination of ECS, which diverges remarkably in GCMs (Bony and Dufresne, 2005;Cess et al., 1990;Vial et al., 2013). This considerable uncertainty in GCMs is mainly due to the shallow clouds (Schneider et al., 2017) (see **Figure 1** by Schneider et al. (2017)). Accordingly, GCMs cannot adequately capture whether shallow clouds contribute to a lower ECS (cooling effect, negative feedback) or a higher ECS (heating effect, positive feedback) (Nuijens and Siebesma, 2019). Especially, the scientists in the atmospheric and climate sciences community have not yet understood:

- how the profile of the wind speed, the interaction between moisture and radiation, and horizontal advection modulates shallow clouds
- what mesoscale aggregation of low clouds indicates for the CRE

in the current and warmer climate (Nuijens and Siebesma, 2019).

The mesoscale organization is a concept widely discussed in cloud studies. Various formations of clouds which are often heterogeneously distributed in the horizontal dimension, ranging from open and close cells in Stratocumulus to cold pool patterns in cumulus clouds are denominated as mesoscale organization (Nuijens and Siebesma, 2019). A vital question is to what extent mesoscale organizations can change due to global warming, and how this feedbacks on CRE (Nuijens and Siebesma, 2019). Further, Stevens et al. (2020) proposed a visual interpretation for the four prominent patterns broadly seen in shallow cumulus cloud fields. Those frequently available patterns are called sugar, gravel, fish, and flower (see Figure 1.1). Correspondingly, Schulz et al. (2021) investigated what meteorological conditions and large-scale motions are in favor of distinct cloud patterns. Besides, various researchers, have proposed quantifiable metrics for cloud organizations, which are objective and interpretable. In this regard, Janssens et al. (2021) collected 21 cloud organization metrics and computed all 21 metrics over 5000 clouds scenes. They found that four principal components of those metrics are amazingly able to capture 82% of all metrics' variance.

This study aims to propose new organization metrics based on a complex network theory approach. A network is a set of points (nodes) connected to each other in pairs by lines (edges) (Newman, 2003b). So far, the previously defined organization metrics are mostly the bulk parameters of the cloud field.

However, network theory-based metrics provide advantageous information that originated from the mutual arrangements between the clouds. In other words, network metrics can potentially quantify how cloud objects interact with and coordinate concerning each other on a cloud field. The idea of considering the cloud field as a network is inspired by Glassmeier and Feingold (2017) who investigated Stratocumulus cloud fields as cellular networks for the first time. Since cumulus clouds are not distributed as homogeneously as Stratocumulus clouds in the horizontal domain, we employ the complex networks to be derived from cumulus clouds. It should be noted that it is the first study in the field of atmospheric sciences that cumulus clouds are investigated as complex networks. In this research, cloud objects are considered geometric complex networks where centroids of cloud objects are defined as the network's vertices. Besides, edges are denoted by different approaches such as removed-link Delaunay triangulation graph and weighted Delaunay graph. Consequently, after proposing new network metrics for cloud organization, we are led to investigate whether a network metric can distinctly encapsulate a characteristic that has not been explored by previously defined organization metrics in the cumulus cloud field.

The research question is can a newly proposed network metric explain a variability that is not captured by the previously defined metrics in cumulus cloud fields?
if **Yes**:

- What does a newly defined network metric distinctly measure for cloud organizations?

if **No**:

- What is the relationship between the network metric and the existing metrics?
- Is it possible for a network metric to replace one of the first four PCs of previously proposed metrics? If yes, it means that this specific network metric is individually able to capture a variability which is captured by a range of previously defined metrics in the organization of cumulus cloud fields. Consequently, one may not need to compute a lot of metrics and its related principal components.

This thesis starts with the description of data in chapter 2. chapter 3 explains how the networks are derived from cloud fields. Then, different network schemes are introduced. In the sequel, various network measures are elaborately discussed. Finally, chapter 3 finishes with the explanation of two statistical methods, multiple linear regression model and principal component analysis. chapter 4 indicates the results of this study and is followed by the associated discussion. Eventually, conclusions are presented in chapter 5.

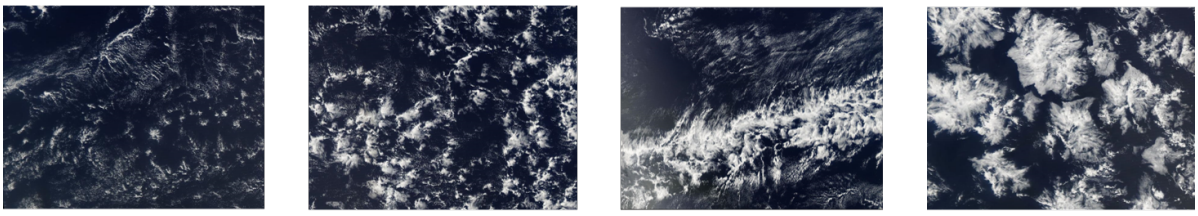


Figure 1.1: The figure reveals the prominent patterns in cumulus cloud scenes. ***Sugar****: A dust of very small-scale clouds that are slightly extended along the vertical direction.* ***Gravel****: A pattern of cloud fields that are along meso-β scale (20-100 km) lines constructing cells. Compared to Sugar, Gravel includes brighter cloud elements.* ***Fish****: Meso-α (200-2000 km) scale clouds in the shape of fish-bone. It is generally associated with a distinct cloud-free region.* ***Flower****: Well-separated cloud objects that are shaped meso-β scale stratiform cloud objects irregularly (Stevens et al., 2020). Figures are taken by Aqua satellite and are from NASA worldview website.*

2

Description of the Data

Similar to studies Stevens et al. (2020); Bony et al. (2020); Janssens et al. (2021), the focus of this project is on the shallow trade-wind cumulus clouds. These cumulus clouds are considerably involved in the monthly and seasonal diversity in cloud fraction, in North Atlantic Ocean, near the east of Barbados (Nuijens and Siebesma, 2019;Nuijens et al., 2014). Cumulus clouds near the east of Barbados ($20^\circ - 30^\circ$ N, $48^\circ - 58^\circ$ W) are appropriately indicative for the whole trades.
The cloud fields are extracted from the Aqua and Terra satellites of the NASA worldview website[1]. The fields are sampled during the daytime from December 2002 to May 2020. It ought to be noted that the data is only downloaded for months from December to May each year.
In this regard, the data-set includes 512×512 km^2 cumulus cloud fields enclosed by $10^\circ \times 10^\circ$ observations. The preprocessing of the data exactly follows Janssens et al. (2021):

1. To increase the size of the data-set, the cloud scenes are let to overlap 256 km.
2. Some of the cloud scenes are removed from the data-set due to the present errors in the productions of remote sensing equipment. To this end, a cloud scene is removed if:
    - Approximately 20% of cloud tops exceeds 5 km
    - The sensor zenith angle is more than 45° (see Figure 2.2)
    - The cloud image is notably affected by the amount of sunlight (see Figure 2.3)

In the next step, the complex networks are derived from the cloud scenes. To analyze complex networks, the network metrics are computed for each cloud field. Those network metrics are elaborately introduced in section 3.3.

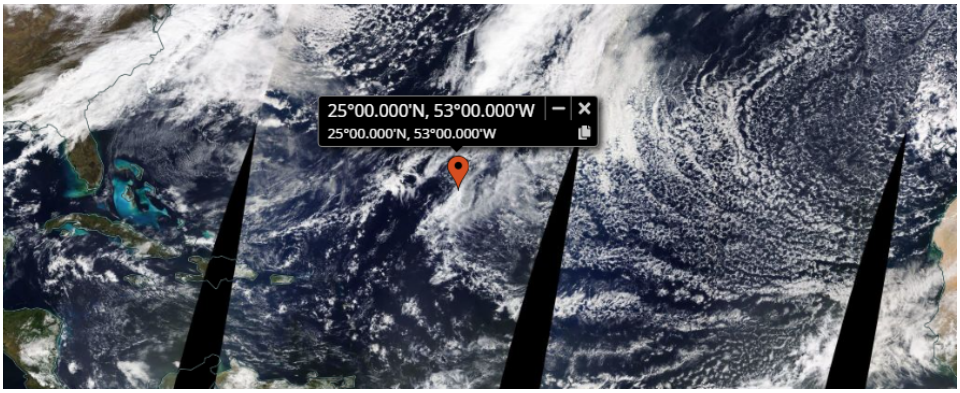


Figure 2.1: The approximate location of interest is shown in the figure.

[1] https://worldview.earthdata.nasa.gov/

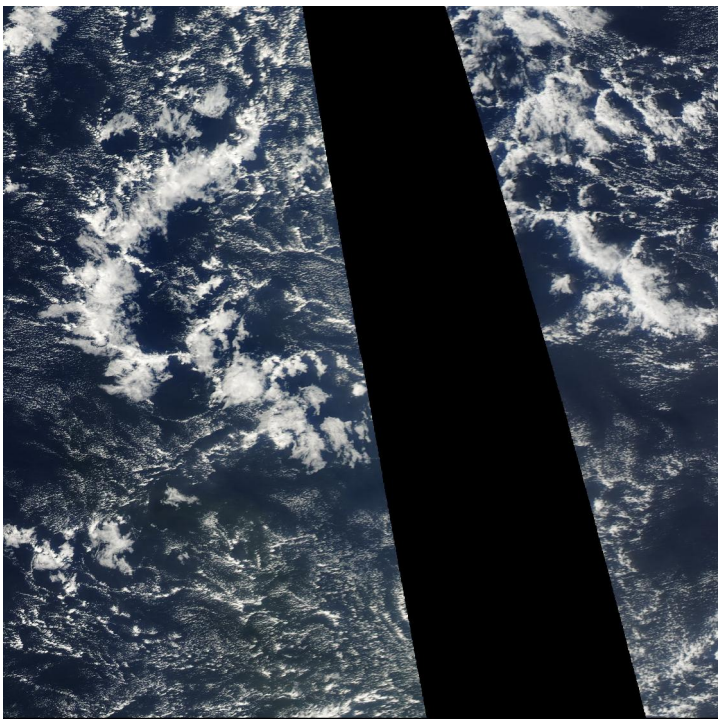


Figure 2.2: An example of a removed scene due to the zenith angle more than 45° and missing data associated with the black region.

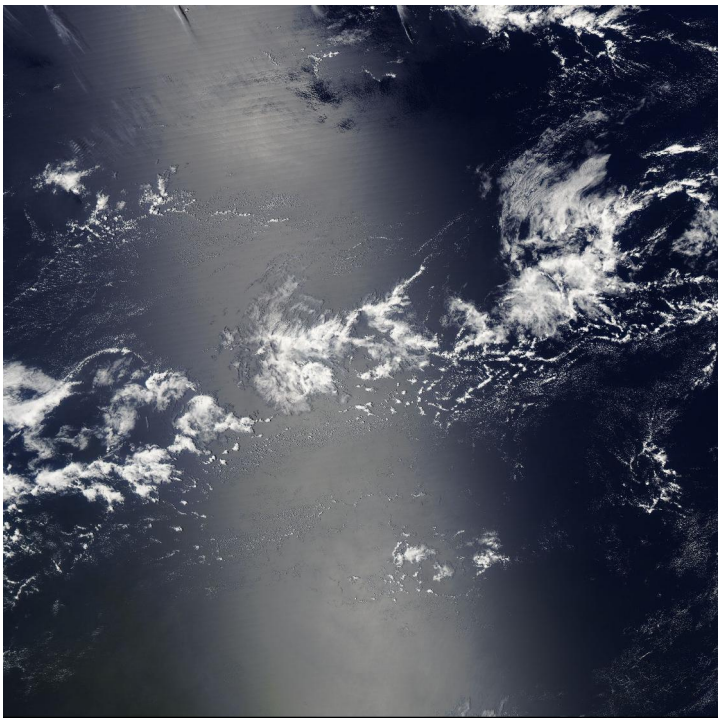


Figure 2.3: An example of a removed cloud field due to the sunlight error. The sunlight is the white thick diagonal shed in the middle of the image. The cloud scenes that suffered from the sunlight error are removed manually from the data set.

3

Methods

3.1. Networks

A network is a collection of objects called *vertices* with links called *edges* that connect vertices. There are numerous systems in the world that shape in the form of networks. Social networks, neural networks, metabolic networks, the World Wide Web, the internet, organizational networks, networks of citations between papers, postal delivery routes, traffic roads, etc., are all real examples of networks in the world. Furthermore, one of the substantial parts of discrete mathematics is the study of the networks using mathematical graph theory. From the first true proof of the network theory by Euler in 1735 to the twentieth century, network theory has expanded up to a vital part of knowledge (Newman, 2003b).

Complex networks have been for long one of the applied branches of mathematics in different fields including computer sciences and social networks. In this regard, complex networks are mostly organized in the shape of graphs where vertices and links are defined in space. Neural networks, Internet, mobile phone networks, transportation systems are all categorized as spatial networks. Spatial networks are graphs where the topology itself does not adequately contain all the characteristics. Rather, space plays a key role. Thus, comprehension of the organization of spatial networks is vital for various fields from epidemiology to urbanism (Barthélemy, 2011). In this research, the aim is to explore the characteristics of cumulus cloud field organizations employing complex spatial networks.

3.2. Networks and Clouds

To derive a network from the cumulus cloud field, the structure of the network must be defined. To this end, the centroids of cloud objects are determined to be vertices of the network. To determine the edges, an important assumption of this research is that each node has a relationship with only its nearest neighbors. In other words, there are edges between each node and only its nearest neighbors. Therefore, all derived networks from cumulus cloud fields are Delaunay graphs (see Figure 3.1). This assumption is firstly made by Glassmeier and Feingold (2017) who derived a Delaunay network from Stratocumulus cloud fields. However, since there is commonly a notable heterogeneity in both of the cloud size distribution and distribution of the distance between the nearest neighbors of cumulus cloud fields, we are led to change the Delaunay graph. Accordingly, two schemes are proposed in subsection 3.2.1 and subsection 3.2.2.

3.2.1. Weighted Delaunay Network

The weighted Delaunay network is the nearest neighbor network with weighted links (see Figure 3.2). To better include the spatial characteristics of cumulus clouds in the network structure, the weight of each link is a function of two important parameters. The first is the area of two connected cloud objects, and the second is the distance between those two objects. In this regard, the link's weight is denoted as:

$$w_{ij} = \frac{\sqrt{A_i} + \sqrt{A_j}}{d_{ij} \cdot \sqrt{\pi}} \tag{3.1}$$

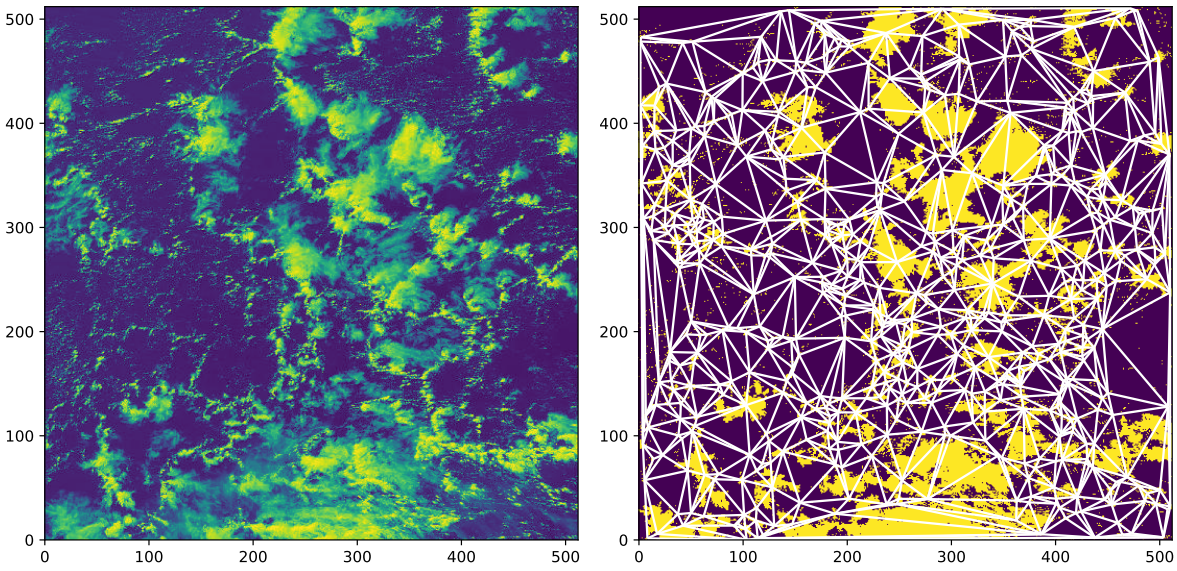


Figure 3.1: Left: Cumulus cloud field from NASA worldview website, right: cloud mask with Delaunay triangulation. Cloud mask is defined as the elements of an image array with values large than 0.5. Each cloud object is an object with an area larger than or equal to 4 pixel-squared. The numbers on the x and y axes are related to the number of image's grids (pixels).

where A is the area of each cloud object and d_{ij} is the distance between the centroids of objects i and j. This idea is inspired by White et al. (2018) who proposed a different cloud organization metric which is called Convective Organization Potential (COP). Considering each cloud object as a circle with the same area and centroid, Equation 3.1 can be written in a different way as:

$$w_{ij} = \frac{1}{d_{ij}} \left(\frac{\sqrt{A_i}}{\sqrt{\pi}} + \frac{\sqrt{A_j}}{\sqrt{\pi}} \right) = \frac{1}{d_{ij}} \left(R_i + R_j \right) \tag{3.2}$$

where, R_i and R_j are the equivalent radii of clouds i and j, respectively. Hence, with respect to the values of the terms R_i and R_j, w_{ij} can obtain these values:

$$w_{ij} \begin{cases} < 1 & \text{if } R_i + R_j < d_{ij} \\ = 1 & \text{if } R_i + R_j = d_{ij} \\ > 1 & \text{if } R_i + R_j > d_{ij} \end{cases}$$

3.2.2. Removed-edge Delaunay Network

In this scheme, some of the edges of the Delaunay graph are removed based on criteria similar to those in subsection 3.2.1. The edge ij of the Delaunay graph is removed if:

- **First criterion.** The mean area of cloud objects i and j is smaller than the 3^{rd} quartile of cloud size distribution of the field,
- **Second criteria.** The distance d_{ij} is larger than the 3^{rd} quartile of the distribution of the distance between all cloud objects in the field.

It should be noted that both conditions need to be fulfilled. Intuitively, it means that there is no relationship between two small cloud objects which are far from each other. An example of a cloud field with the derived removed-edge Delaunay network can be seen in Figure 3.3.

3.3. Network Measures

In case the network's structure is known, it is possible to compute helpful measures that encapsulate the specific characteristics of network topology. Most of the ideas originate from the field of social

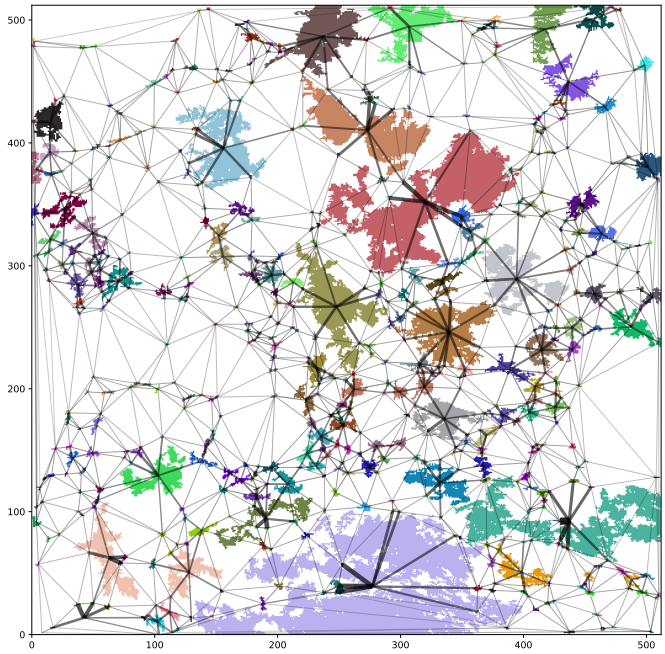


Figure 3.2: The plot represents a cumulus cloud mask with the derived weighted network from it. Cloud objects are painted with different colors, to be better distinguished individually by the human's eyes. Besides, both link's thickness and opacity are proportional to the corresponding weight.

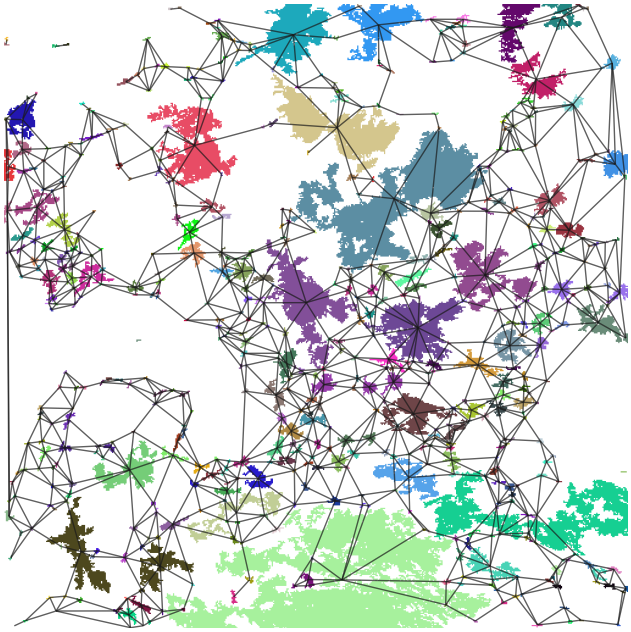


Figure 3.3: The plot demonstrates a cumulus cloud mask with the derived removed-edge Delaunay graph. Compared to Figure 3.1, it is attempted to remove the links that had connected two small objects located far from each other.

network analysis, and much of the terminology used to express these ideas indicate their sociological origin. Nevertheless, the upcoming network measures are widely used in various fields such as biology, computer sciences, physics, and constitute a crucial part of the network toolbox (Newman, 2018).

3.3.1. Degree

In each graph, the degree of node i is equal to the number of edges connected to it (Van Mieghem, 2014). Also, the degree of node i is the sum of the i^{th} row of the graph's adjacency matrix A . For a graph with N nodes, A is a $N \times N$ matrix and it is computed as (Barthélemy, 2011):

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Thus, the degree of node i, k_i is calculated by:

$$k_i = \sum_j A_{ij} \tag{3.4}$$

Similarly, the degree of node i in a weighted graph, is given by Barthélemy (2011):

$$s_i = \sum_{j \in \Gamma(i)} w_{ij} \tag{3.5}$$

where, $\Gamma(i)$ is the set of neighbors of node i, and w_{ij} is the weight that is assigned to the link between nodes i and j.

3.3.2. Clustering

The local clustering coefficient for node i is given by Newman (2018):

$$\mathrm{CC_i} = \frac{T(i)}{\binom{k_i}{2}} \tag{3.6}$$

where $T(i)$ is the number of pairs of neighbors of vertex i that are linked, and the binomial coefficient $\binom{k_i}{2}$ is the number of all possible pairs of neighbors of i. In social networks, the clustering coefficient is explained by the probability that two friends of node i are friends of each other (Newman, 2018).

Clustering coefficient of node i in a weighted graph is defined as Onnela et al. (2005):

$$\frac{1}{k_i(k_i-1)} \sum_{jk} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3} \tag{3.7}$$

where $\hat{w}$ is normalized by the maximum weight in the graph, and $(\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}$ is the geometric mean of assigned weights to the edges of each triangle.

3.3.3. Shortest Path

Generally, there are numerous paths between nodes i and j in a connected graph. The shortest path is written as Barthélemy (2011):

$$l(i,j) = \min_{\text{paths}(i \to j)} |\text{paths}| \tag{3.8}$$

where $|\text{path}|$ is the length of the path and it is given by counting the number of edges along the path. A small average shortest path in the network is associated with the fast spreading of the signals along the network (small-world effect) (Newman, 2018).

The definition of the shortest path in a weighted graph is dependent on how the weights are defined. In this study, the shortest path is given by:

$$l_w(i,j) = \min_{\text{paths}(i \to j)} |\text{path}|_w \tag{3.9}$$

where, if there are N links between nodes i and j, $|\text{path}|_w$ is defined as:

$$|\text{path}|_w = \sum_{n=1}^{N} \frac{1}{w_n} \tag{3.10}$$

Implying that the larger the weights the shorter the path between nodes i and j. This means that if two clouds are larger and closer to each other, the assigned weight to the edge that links them together is larger, thus the path between them is shorter.

3.3.4. Betweenness Centrality

Generally, the term "centrality" of a node in the network reflects the significance of the node. In the network theory community, there are different measures for centrality including closeness centrality, degree centrality, etc. In this study, we selected the betweenness centrality measure. Considering every node pairs s and t, betweenness centrality $g(i)$ of node i is calculated as Barthélemy (2011):

$$g(i) = \sum_{s \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3.11}$$

where σ_{st} is the number of shortest (s,t) paths, and $\sigma_{st}(i)$ is the number of those paths that pass through node i (Brandes, 2008). Intuitively, higher betweenness centrality is an indication of network's resilience to node removal (Newman, 2018;Barthélemy, 2011).

3.3.5. Average Degree of Neighbors

For node i, the average degree of neighbors is given by Barthélemy (2011):

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j \in \Gamma(i)} k_j \tag{3.12}$$

where $\Gamma(i)$ is the set of neighbors of node i.

Similarly, the average degree of neighbors for node i in a weighted network is defined as Barrat et al. (2004):

$$s_{nn}(i) = \frac{1}{s_i} \sum_{j \in \Gamma(i)} s_j \tag{3.13}$$

where s_i and s_j are weighted degrees for nodes i and j.

3.3.6. Assortativity

Assortativity is the inclination of each node to be connected to the other nodes that are similar (or dissimilar) to them in some aspects (Newman, 2003a). The degree of assortativity is determined via the Pearson correlation coefficient of the degree-degree correlation (Allen-Perkins et al., 2017). If the correlation r is between 0 and 1, the network is assortatively mixed. In case $-1 < r < 0$, the network is disassortatively mixed. Besides, if the degree of assortativity is close to 0, the network is a neutrally mixed one (Newman, 2003a). A larger degree of assortativity means high (low) degree vertices are connected to high (low) degree vertices.

3.3.7. Algebraic Connectivity

One of the important branches of mathematics is spectral graph theory that investigates the network's structure through calculating the eigenvalues of the network's Laplacian. In this regard, algebraic connectivity is a measure of how connected a network is and denoted as the second smallest eigenvalue of the network's Laplacian matrix (Newman, 2018). Laplacian matrix of each graph is defined as:

$$L = D - A \tag{3.14}$$

where $D_{ij} = k_i \delta_{ij}$ (δ_{ij}: identity matrix) is the diagonal matrix of vertices' degrees and A is the graph's adjacency matrix (Barthélemy, 2011). Algebraic connectivity for the disconnected network is zero and it reaches its maximum value for the complete graph. Besides, for weighted graphs, the more the density of high-weighted edges, the more algebraic connectivity.

3.4. Statistical Methods

In this section, two statistical methods are introduced in summary. These two methods are utilized within the analyses of this study.

3.4.1. Multivariate Regression Model

Multivariate regression model is a linear regression model with multiple regressors. Assuming $x_1, x_2, ..., x_n$ as independent features, the model is defined as Montgomery et al. (2021):

$$y = \sum_{i=1}^{n} a_i x_i + \epsilon \tag{3.15}$$

where y is a dependent feature, a_i are coefficients, and ϵ indicates the error term. There are important assumptions that should be taken into account:

- The errors are normally distributed, and are uncorrelated.
- There is a negligible co-linearity between regressors.

To check whether the model is able to adequately capture the variation in the target feature, three crucial items should be noticed:

1. R^2. It indicates to what extent the variation in the target value is explained by the model. For instance, a value larger than 0.5 represents that the model can capture more than 50% of the variation in the data.

2. $F\text{-value}$. This parameter is the output of the $F-$test which is employed to test the global significance of the regression. In other words, $F-$test helps to understand whether the model has an overall adequacy or not. The associated hypothesis is:

$$H_0 : a_1 = a_2 = ... = a_n = 0$$
$$H_1 : a_i \neq 0 \text{ for at least one } i$$

   rejection of the null hypothesis (H_0) indicates that the model is adequate enough to capture the variation in the target value y. In this regard, the larger the $F\text{-value}$ the more adequate the model.

3. $p\text{-value}$. This is the output of the $t-$test which is utilized to test the significance of each regressor. After the global adequacy is met the requirement, we are led to ask which regressor is significant to capture the variation in the dependent feature. The corresponding hypothesis is:

$$H_0 : a_i = 0$$
$$H_1 : a_i \neq 0$$

   The regressor x_i can be removed from the model if the null hypothesis H_0 can not be rejected. If the $p\text{-value}$ is smaller than 0.05 for a_i, the associated regressor x_i is significant at 95% level of confidence.

3.4.2. Principal Component Analysis

The principal component analysis is a famous statistical technique widely used for dimensionality reduction performed by projecting the features' vectors onto the first few principal components. With this method, it is possible to achieve a data-set containing much lower dimensions, while explaining a considerable variation of the original data-set. Also, all the principal components are orthogonal, i.e. they are completely uncorrelated (Tipping and Bishop, 1999).

Commonly, to derive principal components one should employ a "standardized linear projection that maximizes the variance in the projected space" (Hotelling, 1933;Tipping and Bishop, 1999). If the original data includes n vectors, principal components are n unit vectors. The i^{th} principal component is the one that is orthogonal to the first $i-1$ unit vectors (i.e. principal components) that maximize the variance of the projected data. The first principal component is the one that captures the largest

variance in the data set compared to the other principal components, the second principal component captures the second greatest variance of the data, and so on. Intuitively, principal component analysis tries to fit an ellipsoid with n dimensions to the data with n vectors. In this case, principal components are the axes of the $n-$dimensional ellipsoid. A particular axis being small implies that the variation along that axis is therefore small.

4

Results and Discussions

As it is mentioned before, network metrics are calculated for each cloud scene. Then, we do a correlation analyses for network metrics and previously defined organization metrics. To answer the research question, the multiple linear regression method is employed to understand whether a linear combination of previously defined metrics are able to capture a considerable variation in newly defined network metrics. If yes, a network metric cannot distinctly capture a variability that is not explored by previously defined metrics in cumulus cloud organizations, and vice versa.

4.1. Weighted Delaunay Network Metrics

The computed metrics for this graph scheme are different statistical moments of degree, clustering coefficient, the average degree of neighbors, weights, betweenness centrality. Besides, algebraic connectivity, average shortest path, and degree of assortativity are calculated as well. Figure 4.1 indicates the correlation heat-map of all calculated network metrics with each other. Large correlation values are between the degree and weights distribution, as well as along and next to the diagonal, different statistical moments of the same metrics are mostly correlated. This is due to the mathematical relationships between some network metrics such as degree and weight that are defined for nodes and links, respectively. Thus, to avoid analyzing similar distributions, it is attempted to reduce the size of the data-set. To this end, we reduced the number of metrics that are tightly correlated with each other, for example, if two metrics are tightly correlated, one of them is removed from further analyses.
Figure 4.2 illustrates the correlation heat-map of reduced weighted network metrics with each other. The final weighted network metrics that are going to be analyzed are:

1. **Gw degree mean**: the mean of nodes' degree distribution
2. **Gw degree std**: the standard deviation of nodes' degree distribution
3. **Gw cc mean**: the mean of nodes' clustering coefficient distribution
4. **Gw avdegnei mean**: the mean of nodes' average degree of neighbors' distribution
5. **Gw avdegnei std**: the standard deviation of nodes' average degree of neighbors' distribution
6. **Gw alg con**: algebraic connectivity of the network
7. **Gw deg assor**: degree of assortativity of the network
8. **Gw btw mean**: the mean of nodes' betweenness centrality distribution
9. **Gw btw max**: the maximum of nodes' betweenness centrality distribution
10. **Gw ave sh p**: the average of all available shortest paths of the network

In subsection A.1.1, for each network metric, cloud scenes are ordered by the network metrics' values. In each figure of subsection A.1.1, the corresponding network metric increases from left to right. Additionally, the outliers of each metric distribution are removed from the low to high plots.

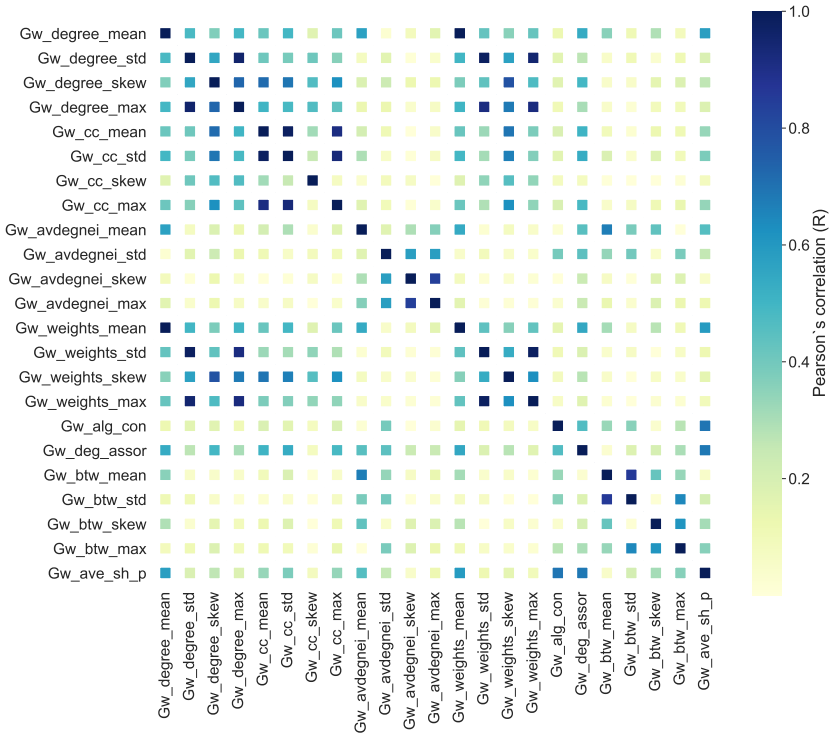


Figure 4.1: Correlation heat-map of all weighted network metrics

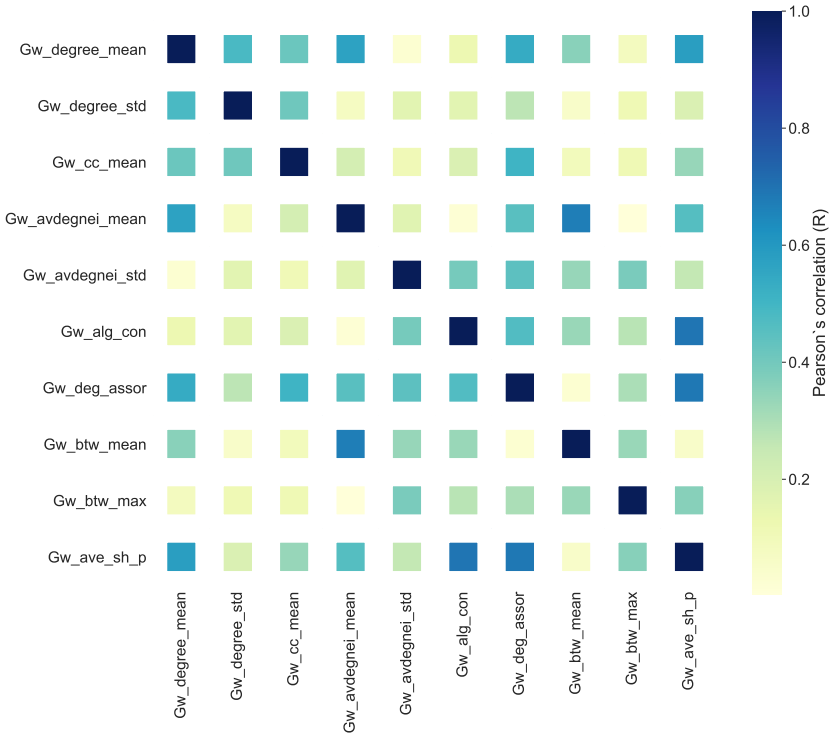


Figure 4.2: Correlation heat-map of reduced weighted network metrics

Figure 4.3 shows the heat-map of correlation values between the weighted network metrics (x-axis) and the previously defined metrics (y-axis). All the metrics illustrated in the y-axis are from the data-set that is prepared by Janssens et al. (2021). Since the first name of the first author is **Martin**, we use this name for the previously defined metrics in the remained parts of the thesis.

Importantly, there are only two correlations with a value larger than 0.7. First is the correlation between the degree of assortativity and size exponent ($|R| = 0.72$), and the second is between fractal dimension and betweenness centrality($|R| = 0.73$). Accordingly, the reason for the tightly correlated metrics is explained in section 4.3. Interestingly, some of the weighted network metrics have low correlation with Martin's metrics. Consequently, section 4.4 investigates what poorly correlated metrics with Martin's metrics distinctly imply for cloud organization.

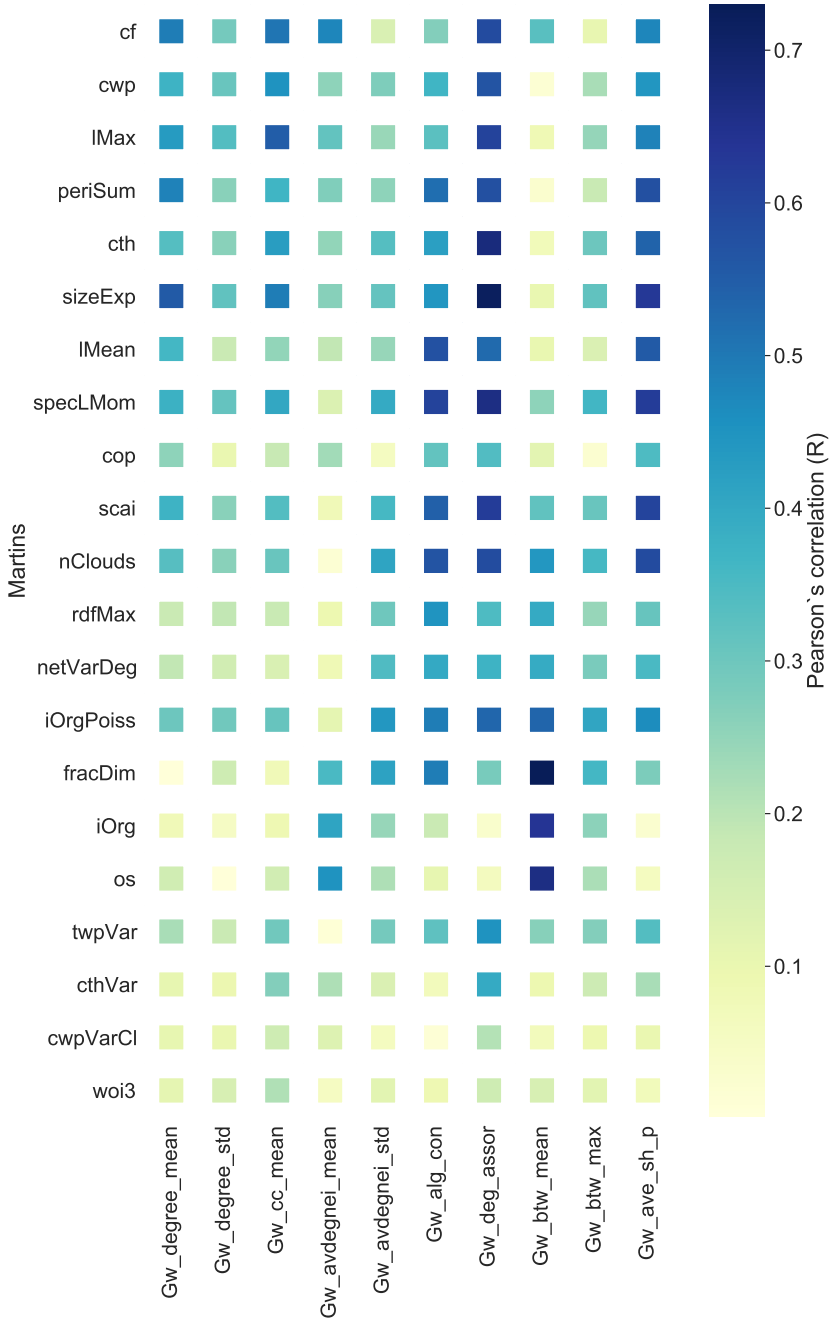


Figure 4.3: Correlation heat-map of network metrics with Martin's metrics. "Martins" refers to Janssens et al. (2021)

4.2. Removed-edge Delaunay Network Metrics

For this graph scheme, the calculated metrics are also various statistical moments of degree, clustering coefficient, the average degree of neighbors, betweenness centrality. In addition to mentioned metrics, the degree of assortativity is computed for the removed-link network scheme. Unlike the weighted scheme, algebraic connectivity and average shortest path are not calculated for the removed-link networks, since the removed-link networks are all disconnected, thus algebraic connectivity and average shortest path are not helpful measures for the analysis of them. Figure 4.4 indicates the correlation heat-map of all calculated network metrics with each other.

In subsection A.1.2, for each network metric, cloud scenes are ordered by the network metrics' values. In each figure of subsection A.1.2, the corresponding network metric increases from left to right. Additionally, the outliers of each metric distribution are removed from the low to high plots.

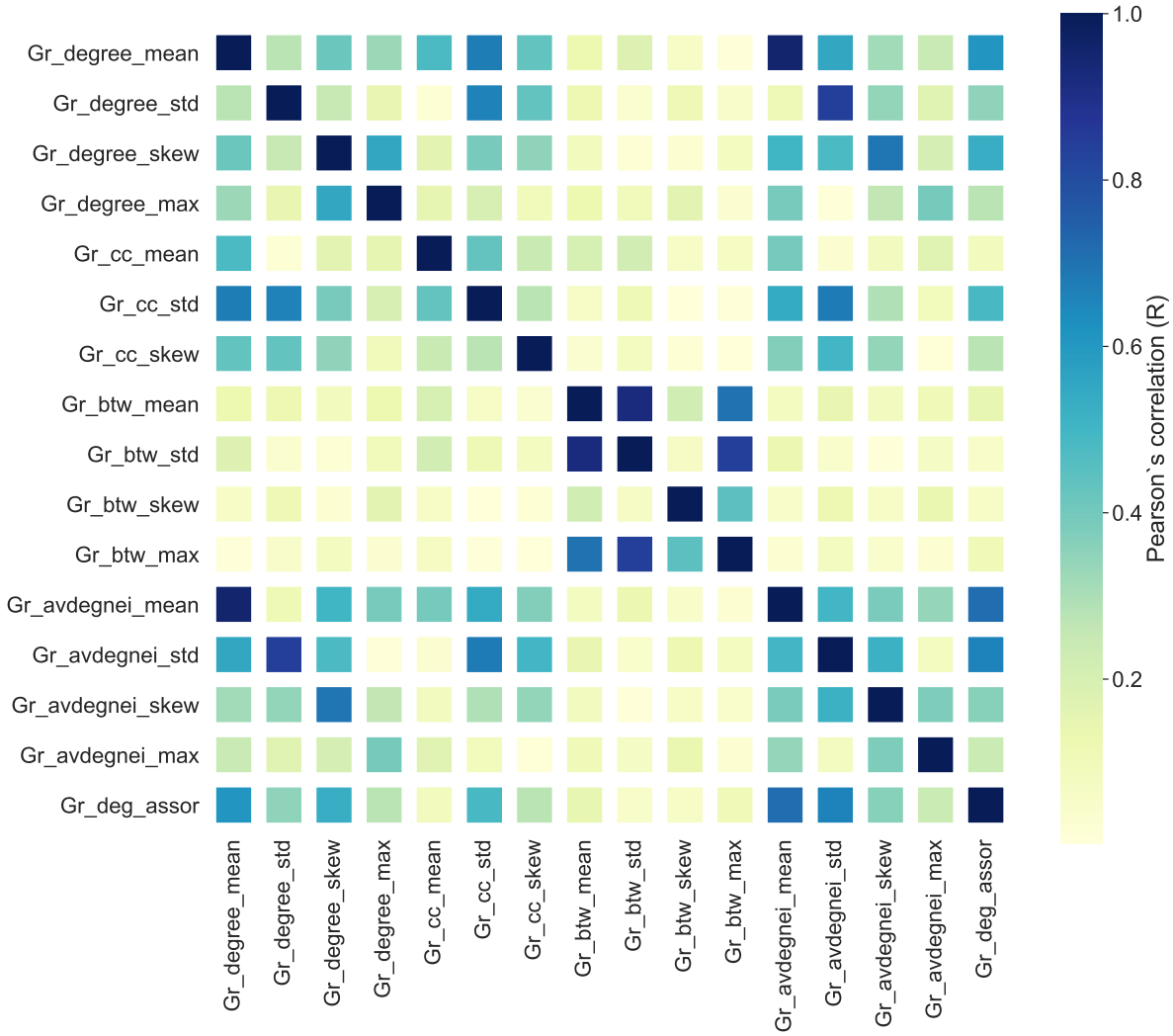


Figure 4.4: Correlation heat-map of all removed-link network metrics

Furthermore, Figure 4.5 depicts how removed-link network metrics are correlated with Martin's metrics. Surprisingly, the maximum of all correlation values is smaller than 0.5. Therefore, it indicates that all removed-link network metrics are completely different metrics. Hence, we are led to ask whether the removed-link metrics are nearly random. Recalling from subsection 3.2.2, this scheme is defined based on arbitrary threshold values which are the function of the size and distance distribution of each field. This raises the question of whether that threshold choice results in approximately random metrics.

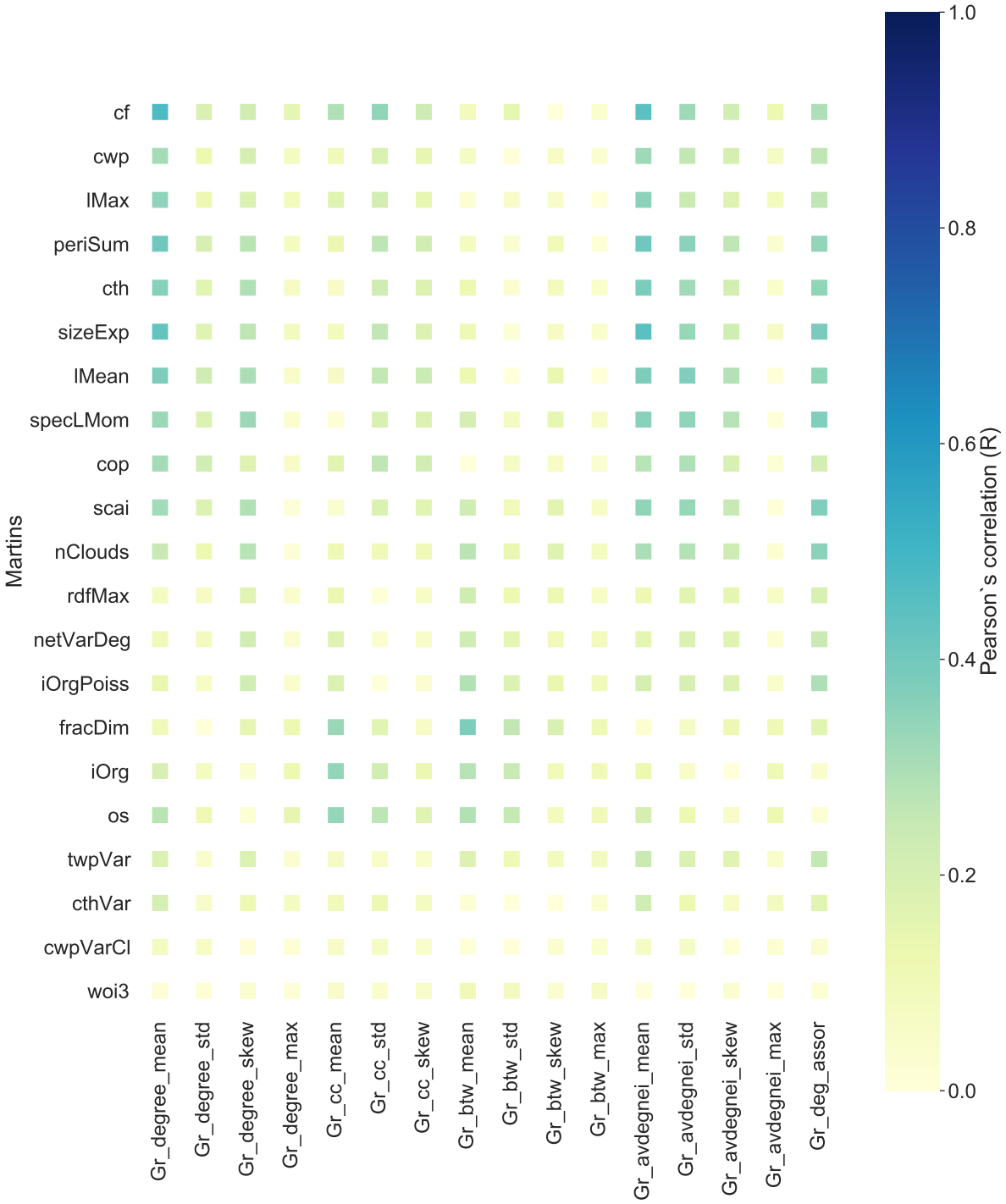


Figure 4.5: Correlation heat-map of removed-edge network metrics with Martin's metrics

To answer the mentioned question on the previous page, an important analysis is performed. To test whether the removed-edge networks are randomly constructed, in addition to removed-edge networks, randomly removed-edge networks are constructed. The removing process for the random version is done based on two conditions:

```
for u, v, weight in G2.edges.data():

    x1 = np.array([0,1])
    r1 = np.random.random(1)
    y1 = x1[0] * (r1 >= 0.5) + x1[1] * (r1 < 0.5)
    x2 = np.array([0, 1])
    r2 = np.random.random(1)
    y2 = x2[0] * (r2 >= 0.5) + x2[1] * (r2 < 0.5)

    if (y1 == 0 and y2 == 0) : G3.remove_edge(u, v)
```

where every two conditions have two different cases, thus, there are four cases in total. In each step, only one case out of all four cases is chosen. In other words, it is attempted to keep everything as much as similar to the removing process based on thresholds, and this makes the analysis somewhat logical. For the comparison, the related code for the removed-link network based on area and distance is as follows:

```
for u, v, weight in G2.edges.data():

    xu, yu = G2.nodes()[u]['position']
    xv, yv = G2.nodes()[v]['position']
    d = np.sqrt((xv - xu) ** 2 + (yu - yv) ** 2)
    area_u = (G2.nodes()[u]['cloud_area'])
    area_v = (G2.nodes()[v]['cloud_area'])

    if d > np.percentile(all_distances, 75) and 0.5 * np.mean(area_u,area_v) <
     ↪  np.percentile(area,75): G3.remove_edge(u, v)
```

To compare the random metrics and removed-edge network metrics, the mean and standard deviation of the distribution of the metrics are computed and plotted versus each other in Figure 4.6. Amazingly, the figure shows that the distribution of both random and removed-edge network metrics are very similar to each other. This result prohibits us to continue the analysis of this scheme in the remained parts of the thesis.

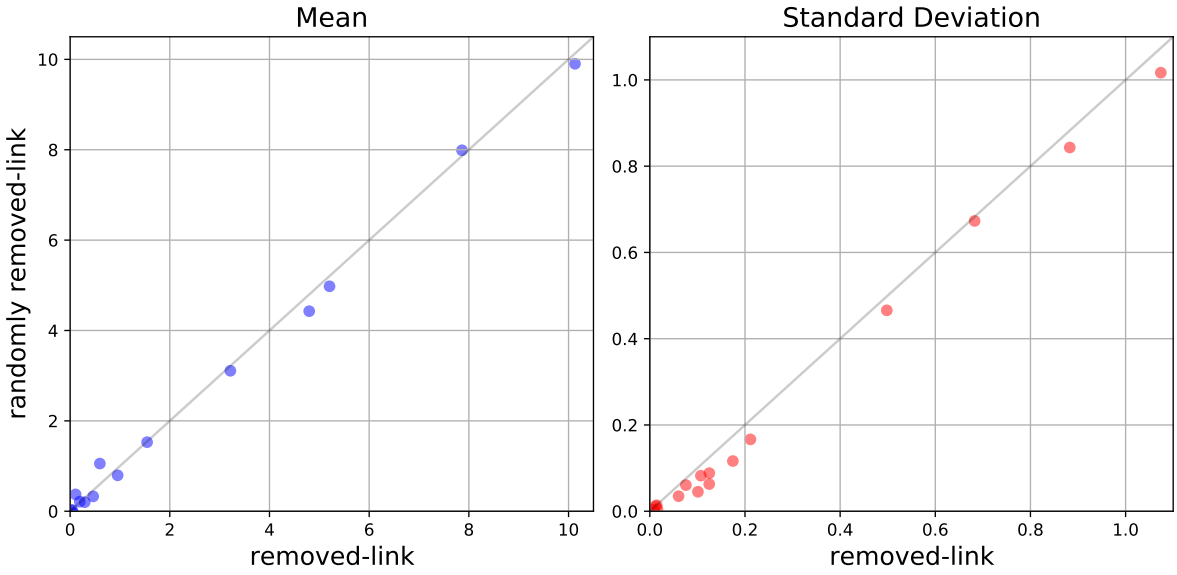


Figure 4.6: The mean (left) and standard deviation (right) values of the distribution of all removed-edge network metrics and random metrics are plotted versus each other. The gray diagonal line is the plot of equation $y = x$.

4.3. Highly Correlated Metrics with Previously Defined Metrics

In this section, how and why the tightly correlated network metrics and Martin's metrics are related is explored. Although it seems challenging to find and understand the justification, an appropriate visualization of results may significantly clarify the interpretation.

4.3.1. Degree of Assortativity & Size Exponent

Figure 4.7 displays the two-dimensional histograms of assortativity degree (y-axis) and size exponent (x-axis). The Pearson correlation between these metrics is negative with an absolute value of 0.72 i.e., with increasing the assortativity degree, the value of size exponent is decreasing. Additionally, the degree of assortativity distribution manifests that this metric for the cloud fields is limited between 0 and 0.8, implying that all cloud fields are assortatively mixed networks.

The size exponent b is obtained by

$$\log N_c \propto b \log l \tag{4.1}$$

where N_c is the number of all clouds with sizes in bins that their width are increasing exponentially. The term l is a length scale that is calculated as $l_i = \sqrt{A_i}$ for the cloud object i with the area A_i (Ding et al., 2014;Janssens et al., 2021).

Recalling from subsection 3.3.6, higher values of assortativity degree is associated with a larger tendency of larger (smaller) clouds to be connected with larger (smaller) clouds. Figure 4.8 shows that in cloud fields with a larger degree of assortativity, small clouds mostly tend to be linked to small clouds. As figure 4.8a shows, this is related to a cloud field with a more homogeneous cloud size distribution. Inversely, Figure 4.9 indicates that the more the homogeneity of cloud size distribution, the more the value of size exponent. Hence, the larger the degree of assortativity, the smaller (larger) the value (absolute value) of the size exponent.

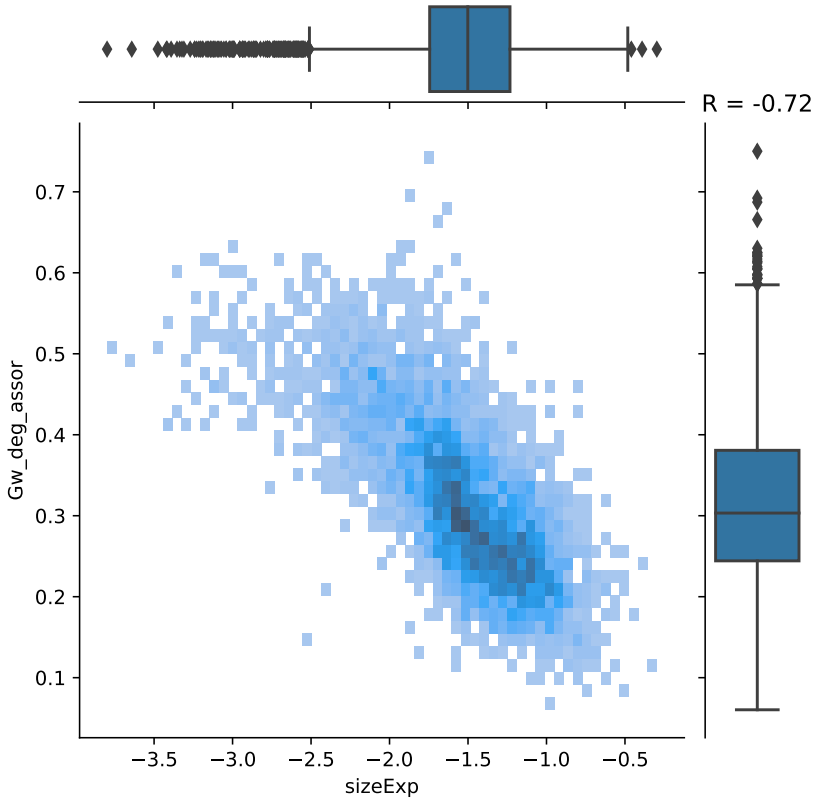
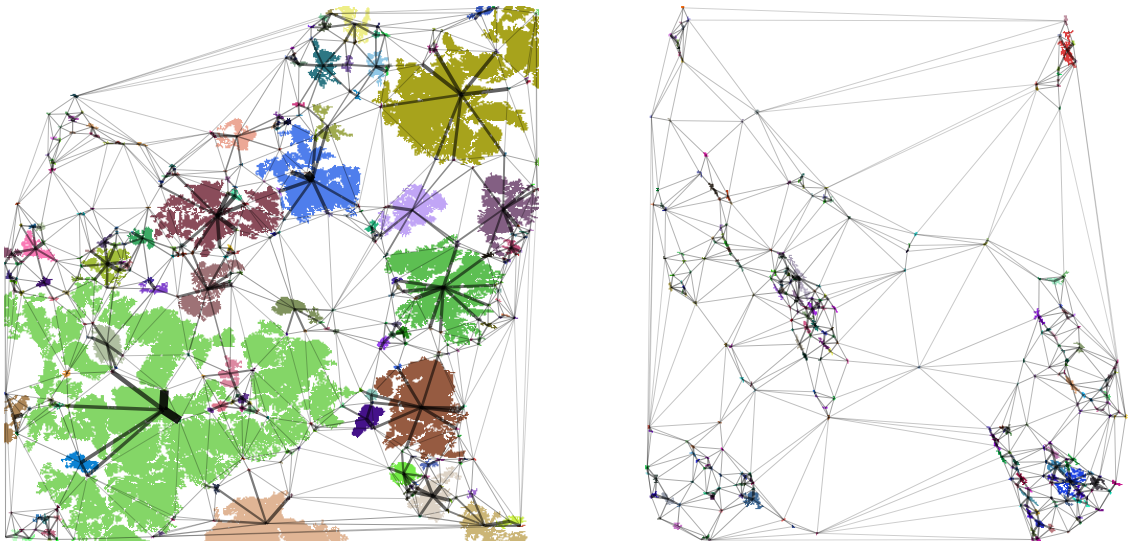
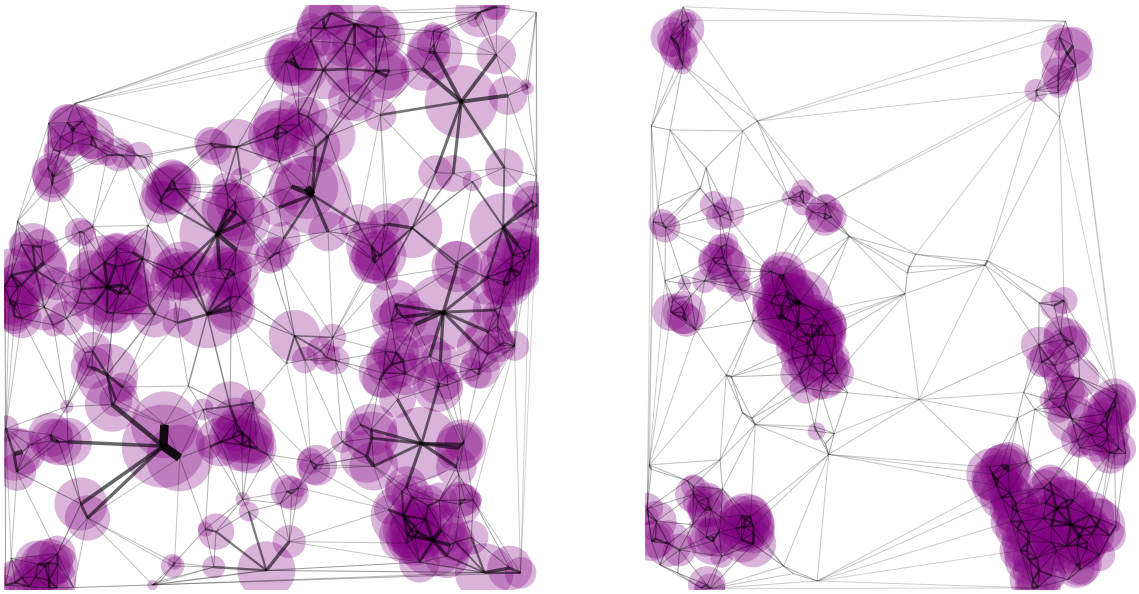


Figure 4.7: The figure shows the two-dimensional histograms of degree of assortativity and size exponent.

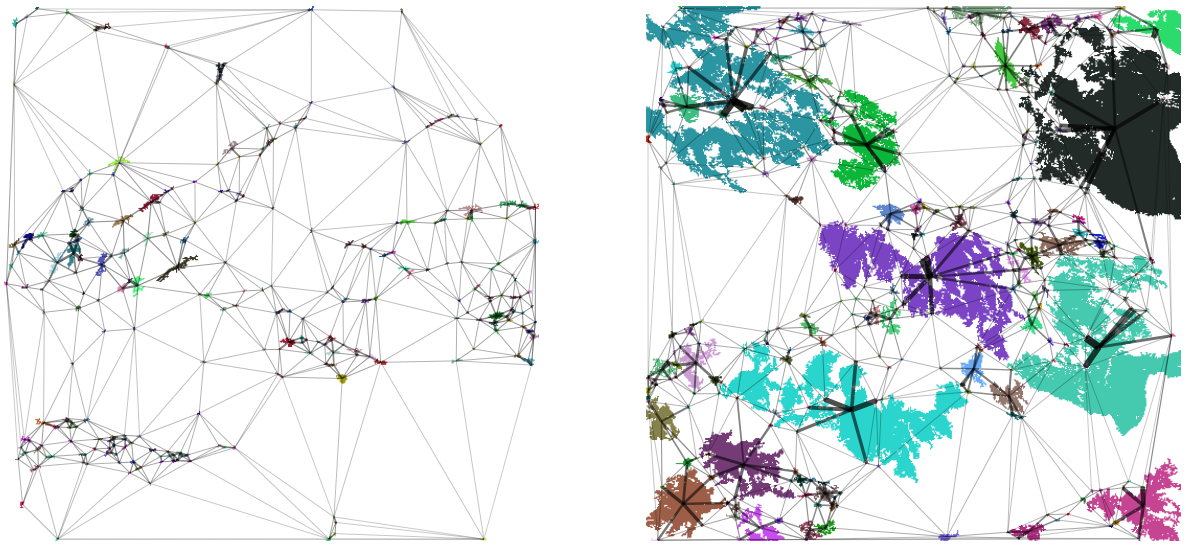


(a) Clouds

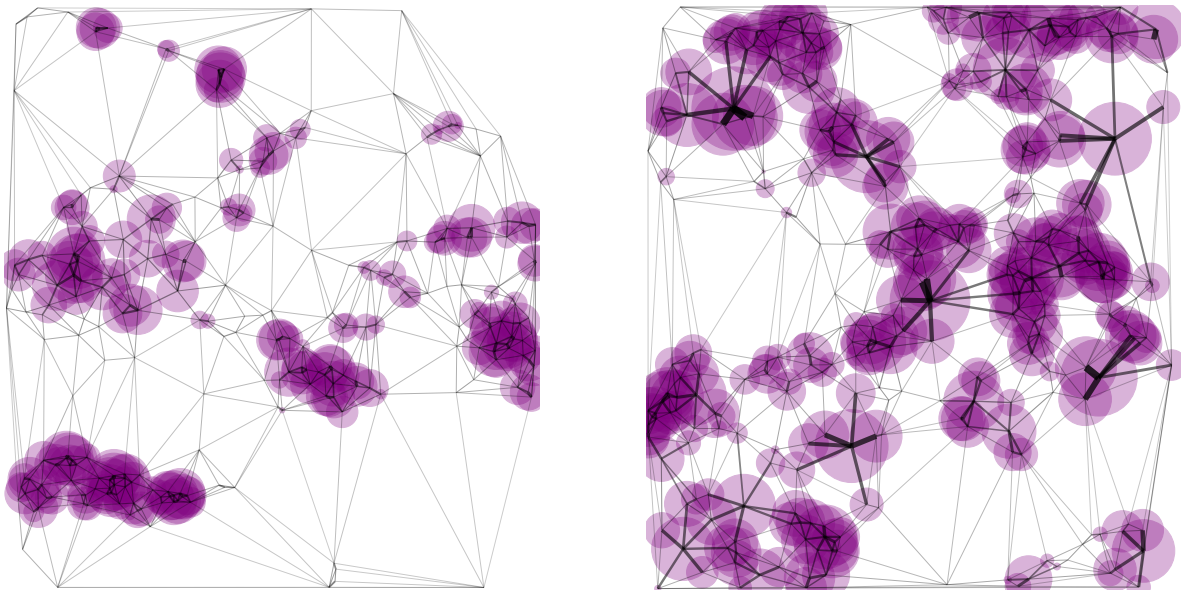


(b) Network; the size of each circle is scaled by the node's degree.

Figure 4.8: The plots are arranged from low to high (left to right) by the degree of assortativity.



(a) Cloud scenes



(b) Networks; the size of each circle is scaled by the node's degree.

Figure 4.9: The value of size exponent increases from left to right.

4.3.2. Betweenness Centrality & Fractal Dimension

Two-dimensional histogram of betweenness centrality mean (y-axis) and the fractal dimension (x-axis) is illustrated in Figure 4.10. The two-dimensional histogram explains a negative correlation between the metrics i.e., for a cloud field, the larger the fractal dimension, the lower the mean betweenness centrality.

The fractal dimension D_f is derived by

$$\log N_c = D_f \log l_b \tag{4.2}$$

where N_c is the number of square boxes with size l_b that are neither completely filled by clouds nor completely cloudless. The term D_f is computed as a slope of fitted line to Equation 4.2 (Cahalan and Joseph, 1989;Janssens et al., 2021).

The investigation of the relationship between the fractal dimension and mean betweenness centrality appears to be complicated though, visualization helps us to intuitively comprehend the association of these two metrics. In this regard, larger values of fractal dimension are corresponded to sugar types of clouds (see Figure 4.11), accordingly the fields covered by the sugar clouds poss smaller values of mean betweenness centrality (see Figure 4.12). In other words, a cloud field with homogeneously distributed small cloud objects, weights are relatively smaller compared to the other fields, consequently, this reduces the availability of nodes with relatively high betweenness centrality.

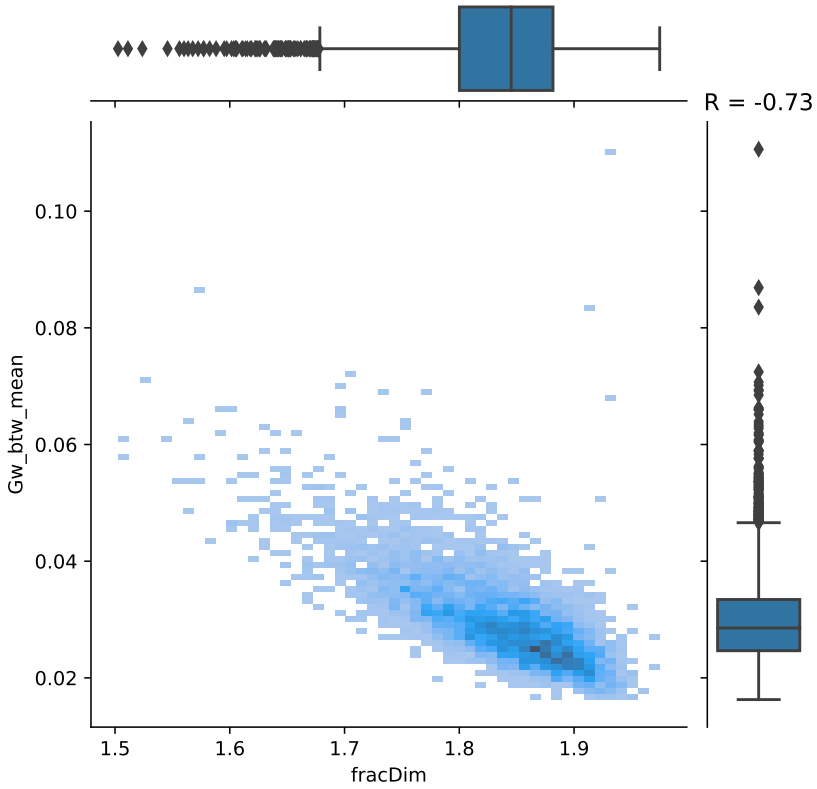
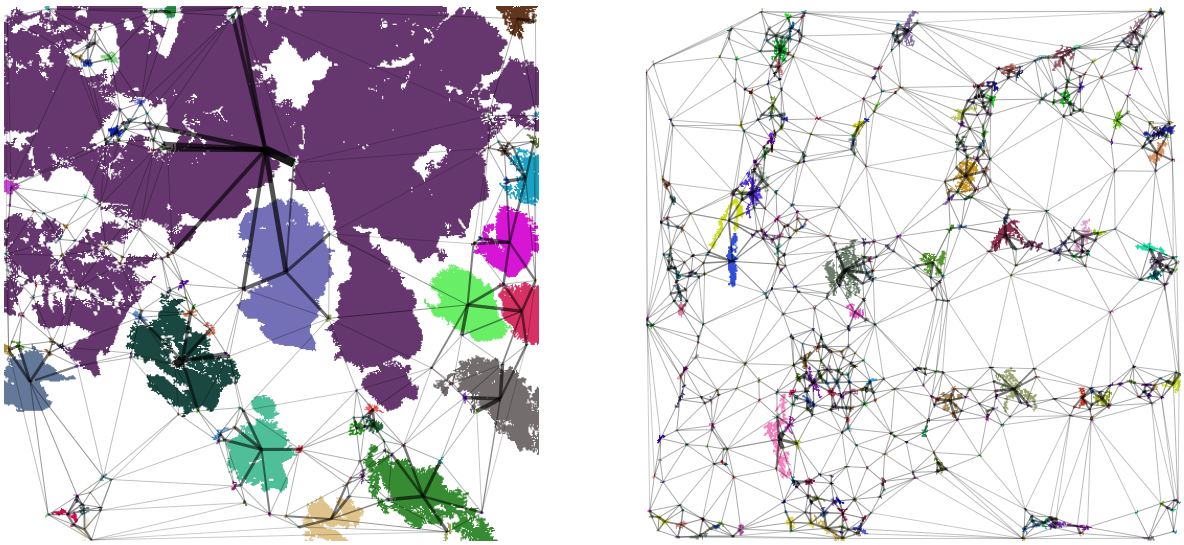
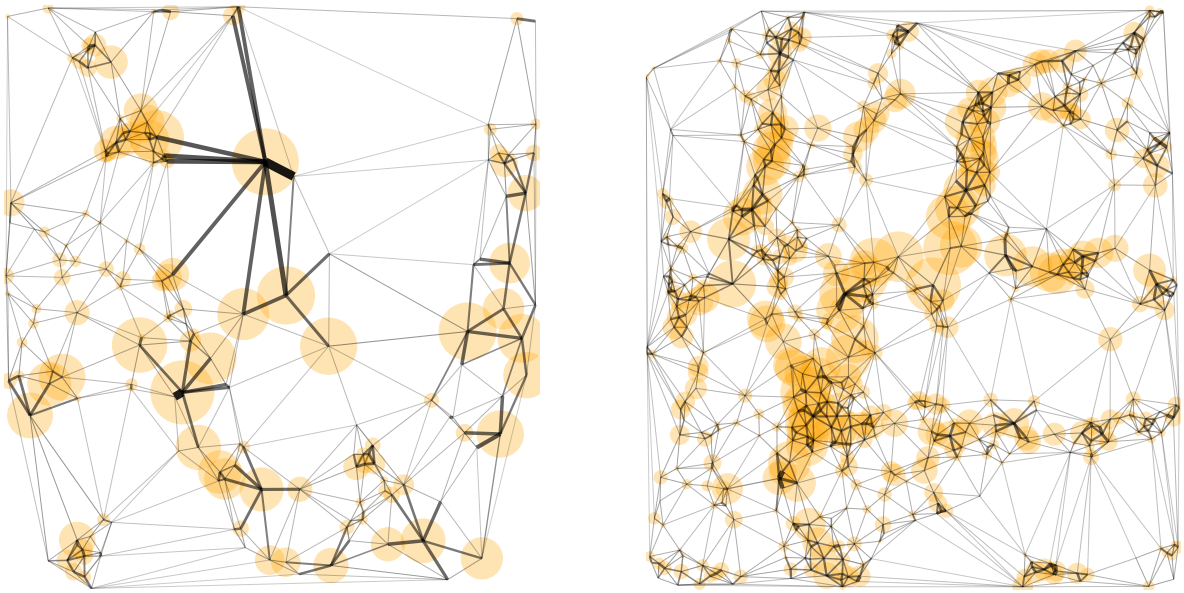


Figure 4.10: The figure shows the two-dimensional histograms of mean betweenness centrality and fractal dimension.

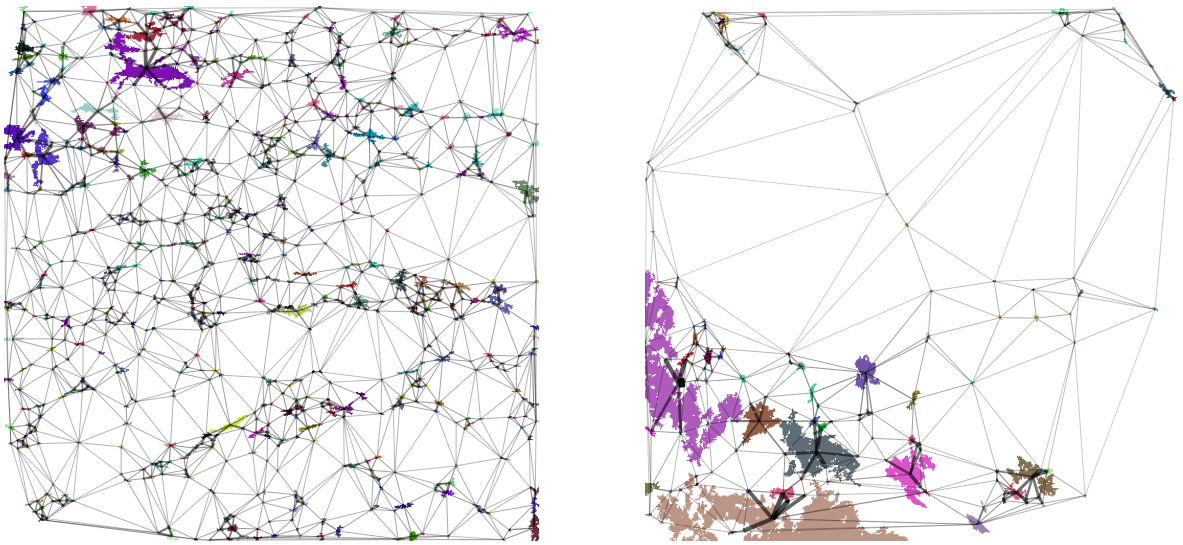


(a) cloud

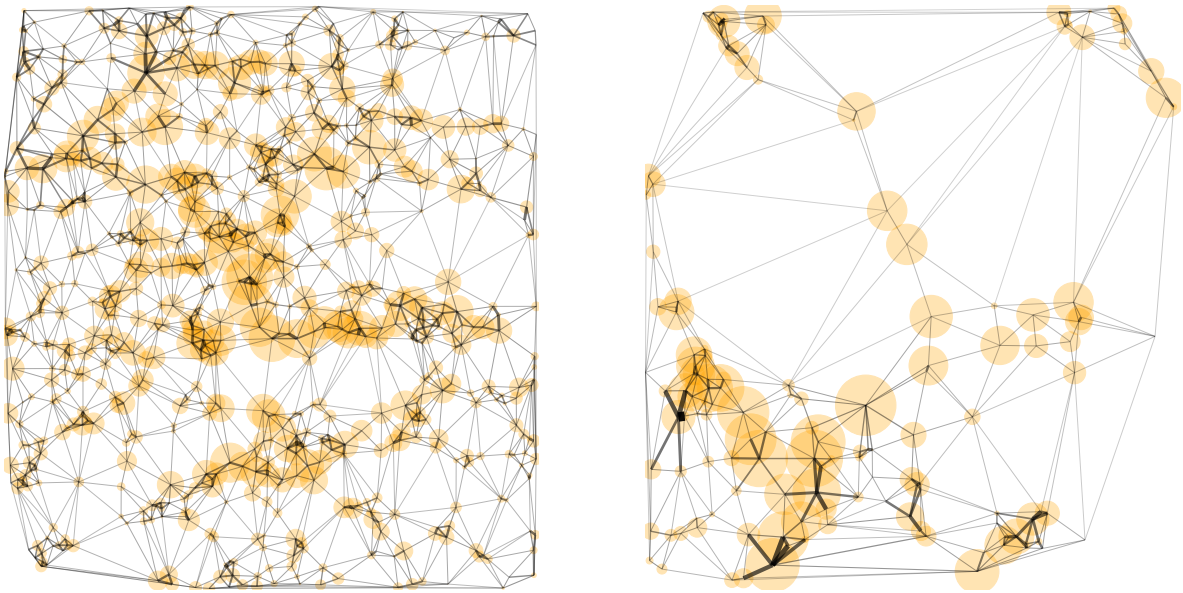


(b) Network; the circle sizes are scaled by the nodes' betweenness centrality.

Figure 4.11: Value of fractal dimension grows from left to right.



(a) cloud



(b) Network; the circle sizes are scaled by the nodes' betweenness centrality.

Figure 4.12: The scenes are arranged by the value of mean betweenness centrality from low (left) to high (right).

4.3.3. Can A Network Metric Replace One of the Principal Components?

One of the research questions of this study is whether a network metric can replace one of the first principal components of the existing metrics. Figure 4.13 illustrates that the maximum absolute value of Pearson's correlation between the network metrics and the first four principal components of Martin's metrics is 0.72. **Figure S3.** of Janssens et al. (2021) represents that the size exponent and the fractal dimension have correlation values of more than 0.8 and 0.7 with the first and second principal components, respectively. The absolute correlation values of assortativity degree and average betweenness centrality are 0.72 and 0.68 with the first and second principal components. In addition to what is mentioned, considering Figure 4.3 and Figure 4.13 led us to deduce that none of the network metrics can replace one of the first principal components of Martin's metrics, since Figure 4.3 shows that the maximum absolute value of correlation between network metrics and Martin's metrics is 0.73.

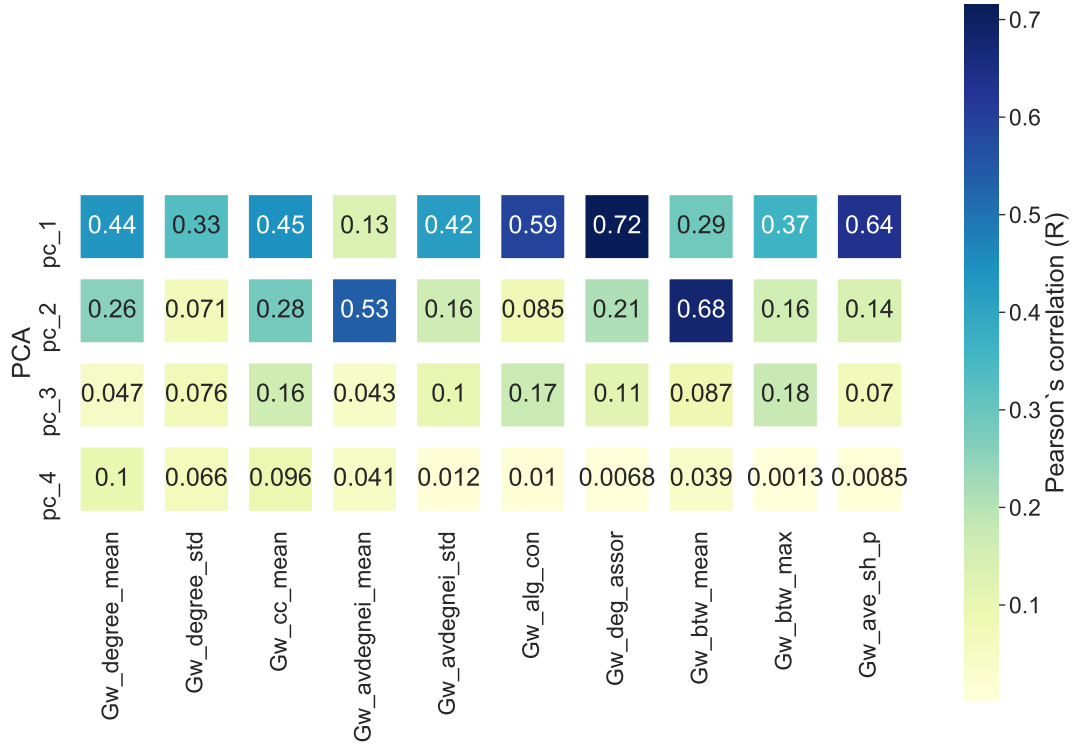


Figure 4.13: Correlation heat-map of weighted network metrics with the first four principal components of Martin's metrics

4.4. Poorly Correlated Metrics with Previously Defined Metrics

In this section, the aim is to probe what distinctly the poorly correlated weighted network metrics imply for the cloud organization. Firstly, Figure 4.13 represents the correlation heat-map between the weighted network metrics and the first four principal components of Martin's metrics. The degree of assortativity and betweenness centrality are tightly correlated with the first and second principal components, respectively. On the other hand, Figure 4.13 reports relatively low values of correlation between the principal components and other weighted network metrics.

Furthermore, we are led to answer the question of whether a linear combination of Martin's metrics can capture the variation in each network metric. To this end, a multiple linear regression analysis is done as:

$$y_j = \sum_{i=0}^{20} a_i x_i + \epsilon$$

where x_i (each regressor) is one of the principal components of Martin's metrics, a_i are the coefficients of the model, and $y_j (j = 0, 1, ..., 9)$ is one of the network metrics. Thus, there are 10 models in total. Correspondingly, Figure 4.14 indicates that the associated model cannot notably capture the variation of degree standard deviation, mean of clustering coefficient, the standard deviation of the average degree of neighbors, and the maximum betweenness centrality. Meaning that the aforementioned metrics can explain a variation in cumulus cloud organization which is not explained by the previously defined metrics.

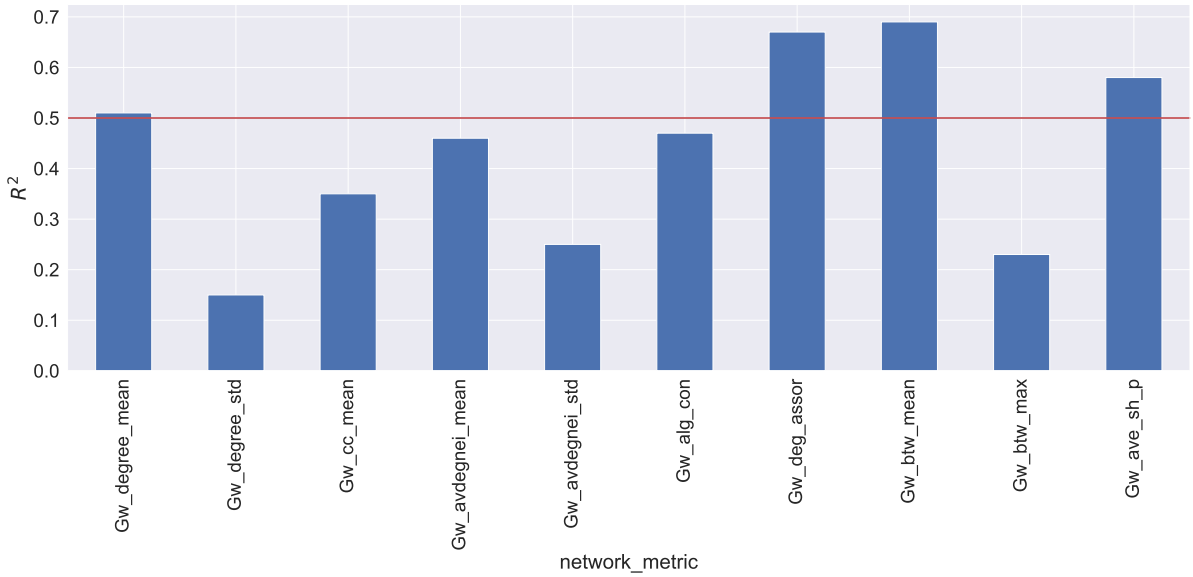
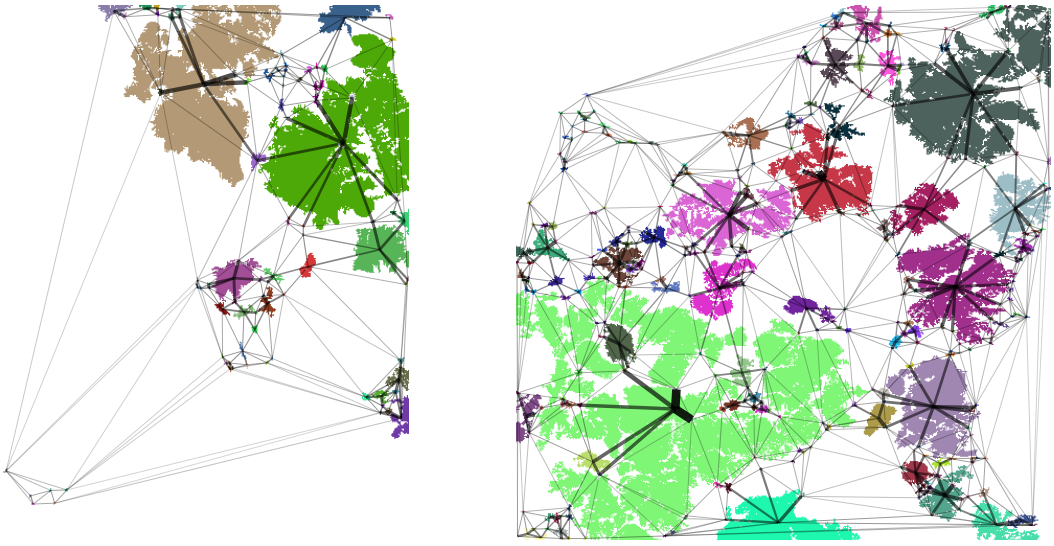
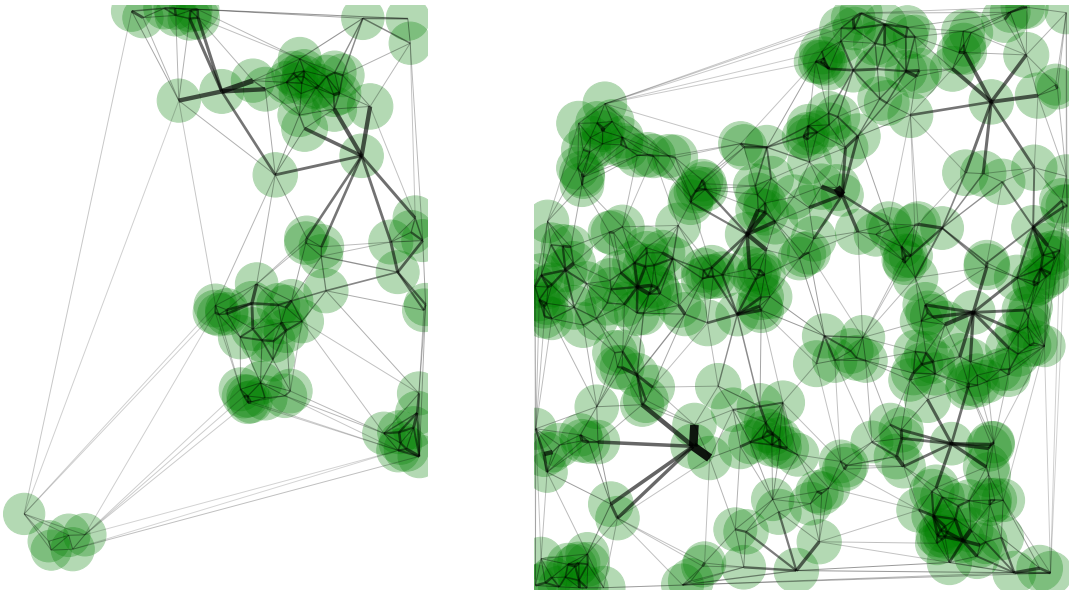


Figure 4.14: The figure shows the R-squared values of each multiple linear regression model. In each model, the target feature is one of the network metrics and the regressors are all 21 principal components of Martin's metrics. To recall, the name of weighted network metrics can be found under section 4.1.

Before starting to see how strikingly the aforementioned metrics explain a variation in cloud fields, the author wants to not further analyze the average degree of neighbors' standard deviation and the maximum betweenness centrality. As Figure 4.15 shows, the average degree of neighbors is approximately the same value for each node. This exactly matches with Equation 3.12 implying that the effect of weight is somehow removed from the calculation of neighbors' degree mean. Also, Figure 4.16 illustrates that the maximum betweenness centrality is not an appropriate proxy of each field, since the maximum value of each distribution can be extremely out of the bulk of the corresponding data-set. subsection 4.4.1 and subsection 4.4.2 investigate the metrics degree standard deviation and clustering coefficient mean in detail.

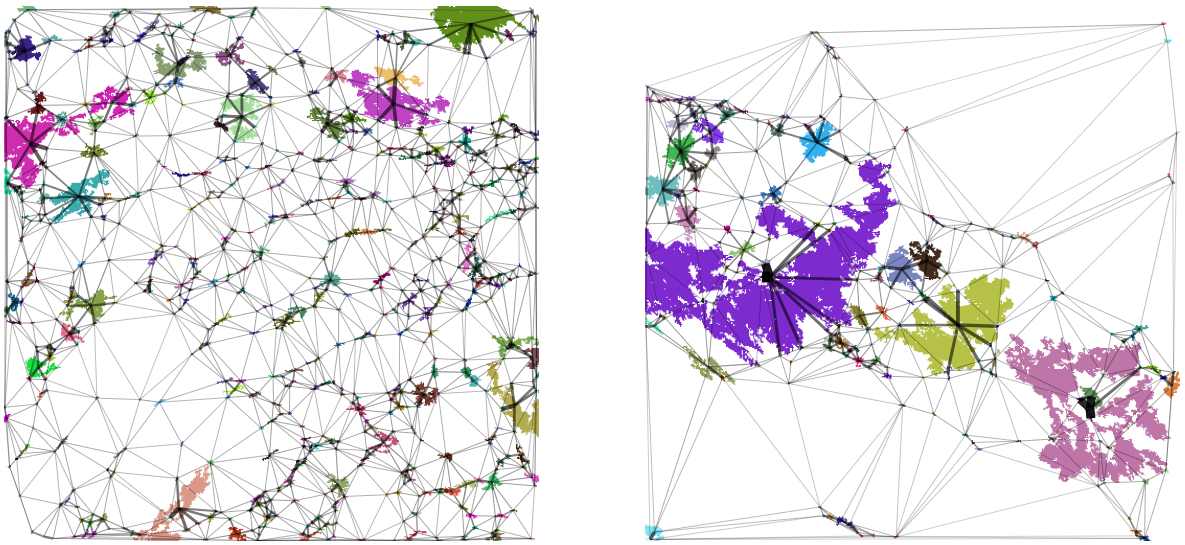


(a) cloud

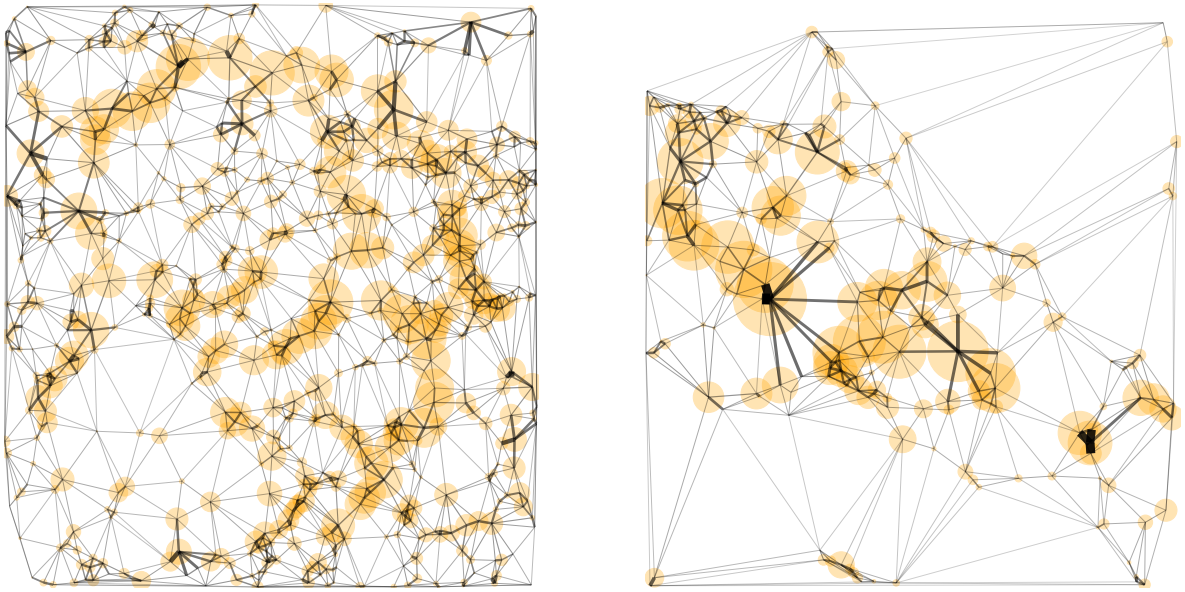


(b) Network; the circle sizes are scaled by the average degree of neighbors' node.

Figure 4.15: The average degree of neighbors gets larger from left to right.



(a) cloud



(b) Network; the circle sizes are scaled by the nodes' betweenness centrality.

Figure 4.16: The maximum betweenness centrality becomes larger from left to right.

To address the goal of this section which is to explore why and how differently a network metric implies for the cloud organization, it is attempted to limit the first four principal components of Martin's metrics in central bins i.e., $-0.5 < PC_i < 0.5, i = 1, 2, 3, 4$ (see Figure 4.17). By limiting the principal components in central bins (red squares in Figure 4.17), they are somewhat prohibited to change considerably. Hence, it makes the analysis possible to see how tangible a network metric explains a variation in cumulus cloud scenes. To this end, in subsection 4.4.1 and subsection 4.4.2 the cloud scenes are ordered from low to high in central bins.

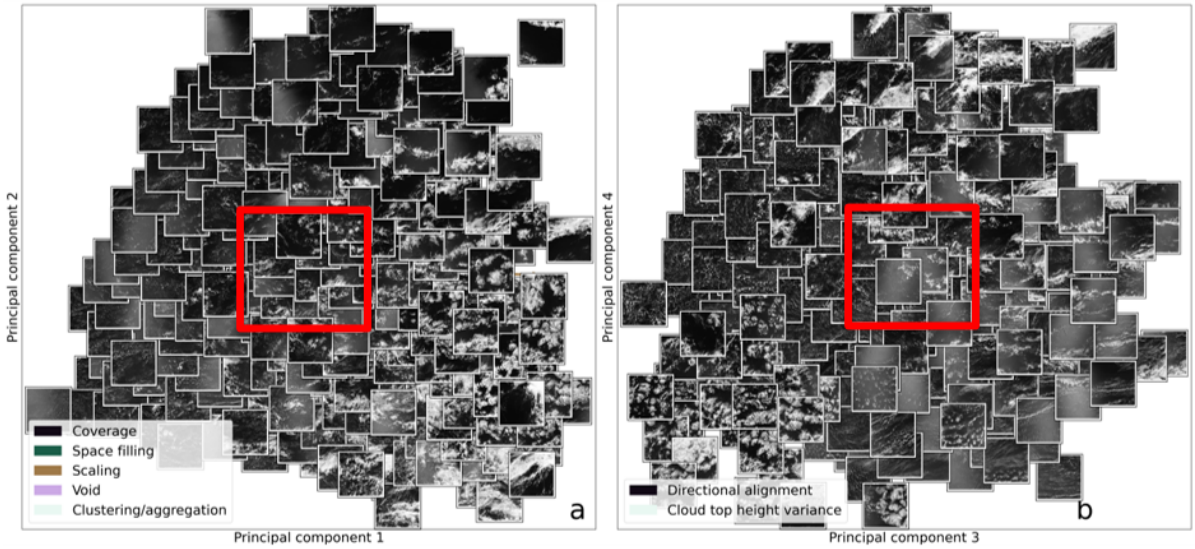
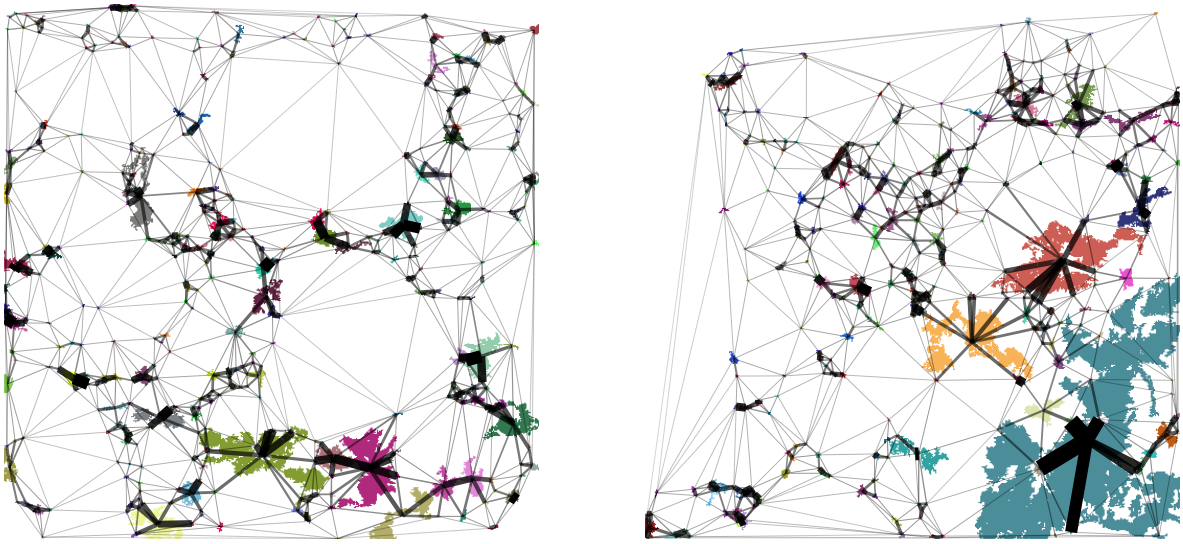


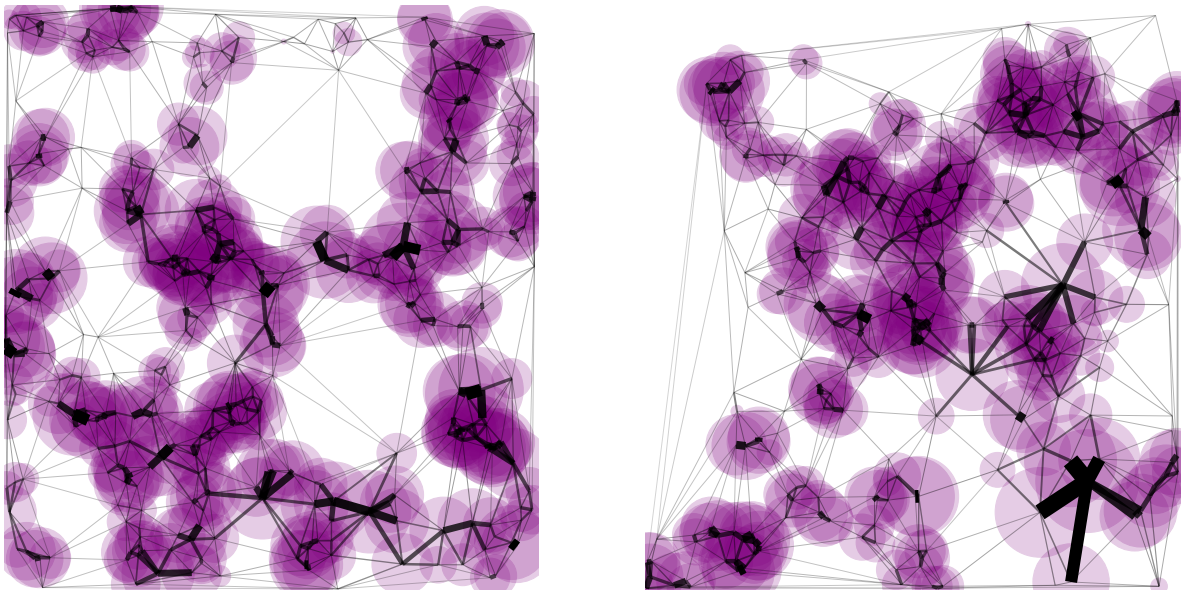
Figure 4.17: The figure shows the cloud fields scattered on a plane spanned by the first four principal components (PCs) of Martin's metrics. The figure is originated from the code and data of Janssens et al. (2021). The red squares represent the central bins approximately. In central bins, there are only 19 out of 5000 cloud scenes.

4.4.1. Degree Standard Deviation

The lower the degree standard deviation, the more homogeneously distributed clouds on the scene i.e., clouds are homogeneously positioned concerning each other, and simultaneously the cloud size distribution is homogeneous (see Figure 4.18). Accordingly, a cloud field with a smaller degree standard deviation potentially contains one organization mode. In contrast, a field with a larger degree standard deviation has a higher potential to be made up of different organization modes. For instance, a field covered by the sugar clouds has a considerably lower degree standard deviation compared to a field that is covered by fish clouds and gravel clouds.



(a) cloud



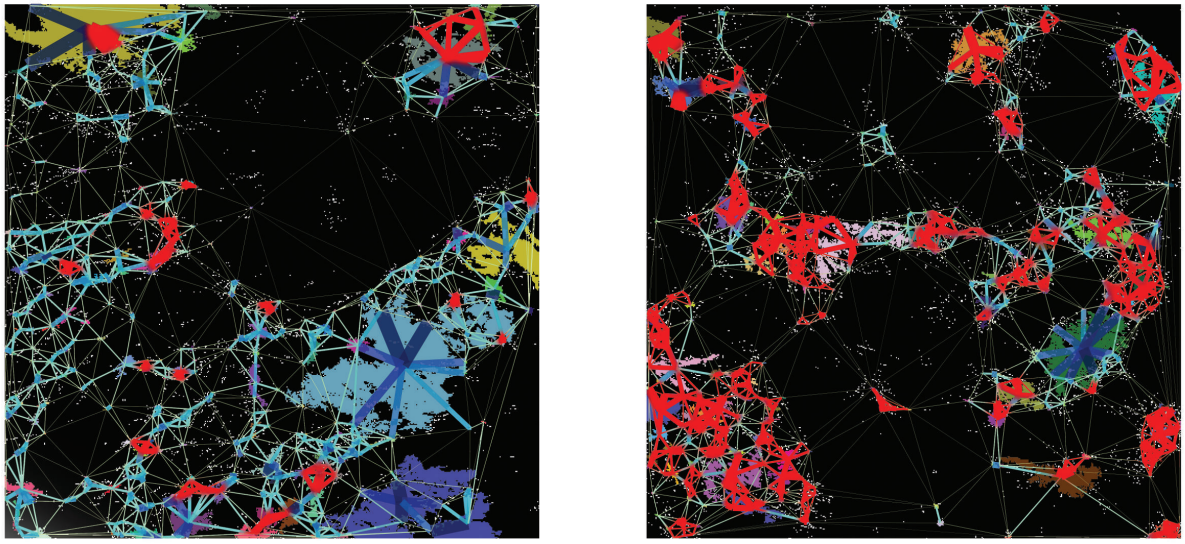
(b) Network; the circle sizes are scaled by the nodes' degree.

Figure 4.18: The cloud scenes are ordered by the degree standard deviation from left to right in central bins.

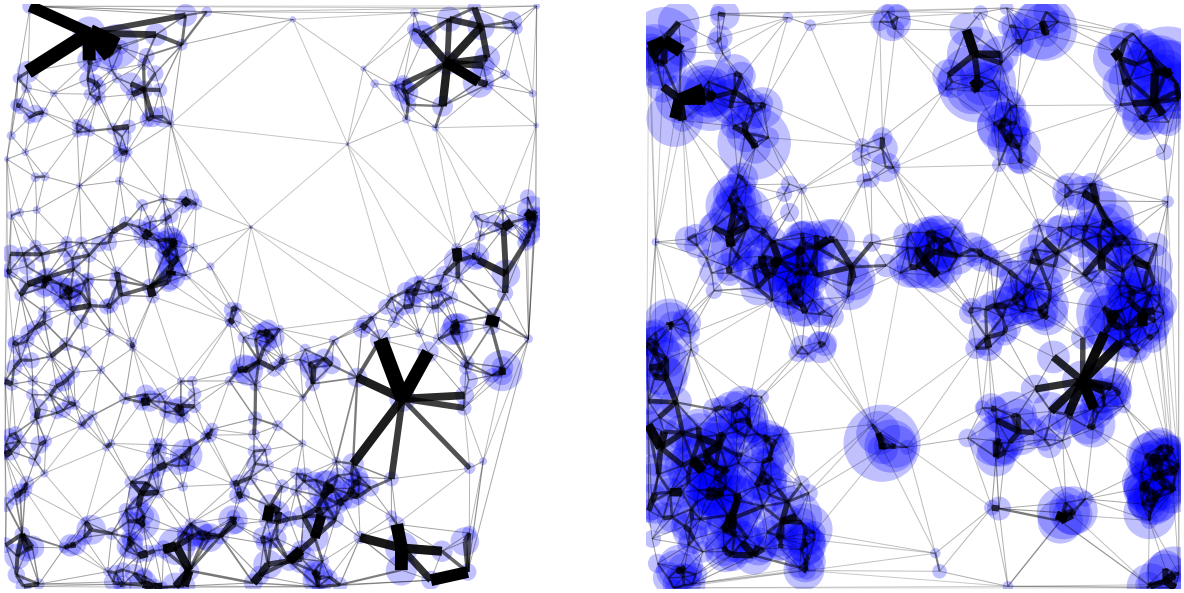
4.4.2. Clustering Coefficient Mean

Figure 4.19 demonstrates that clustering coefficient mean notably changes in central bins. In a Delaunay graph without weighted links, all nodes have relatively the same clustering coefficient. Thus, in the weighted Delaunay network, weights play an indispensable role. Therefore, a cloud field with a large average clustering coefficient is constructed of triangles with high-weighted sides. This can be three cases:

- Relatively larger clouds as a node of each triangle
- Relatively smaller clouds but smaller distances between them
- It can be the combination of the first two cases



(a) Cloud; With increasing the weight of each edge the associated color changes from yellow to blue. Also, the thickness of each link is scaled by the related weight. Besides, It is tried to put a red color for triangles with high-weighted sides (links). It can be seen that the number of strong (red) triangles is notably more in the right plot compared to the left one.



(b) Network; the circle sizes are scaled by the nodes' clustering coefficient.

Figure 4.19: The cloud scenes are ordered by the clustering coefficient mean from left to right in central bins.

4.5. Sensitivity Analyses

4.5.1. Sensitivity to the Distance Term

Recalling the weight function equation in subsection 3.2.1, the d term in Equation 3.1 is the distance between the centroids of cloud objects i and j (center-spacing distance). One could claim that when one cloud object is considerably large, the d term is inevitably large and consequently the weight of the incident link to the associated node is small. This led us to change the d parameter in Equation 3.1 from a center-spacing distance to the edge-spacing distance. In a PhD thesis Laar (2019), Thirza van Laar defined the edge-spacing distance as $d - R_i - R_j$, where d is the center-spacing distance and R_i, R_j are the radii of the circles with the same area and centroid of clouds i, j. Since the cumulus clouds are not perfect circles, this definition may result in negative values of edge-spacing distance. Therefore, in this sensitivity analysis, the definition of edge-spacing distance is slightly different compared to what Thirza van Laar has proposed in Laar (2019). To compute the edge-spacing distance, these steps should be performed:

1. Consider the line that connects the centroids of clouds i, j
2. Compute the pixel distance between the centroids
3. Subtract the cloudy pixels from the total pixel distance

In other words, the edge-spacing distance is the cloudless part of centre-spacing distance.
We re-performed the analysis over the whole data-set to compare the results of both perspectives, center-spacing, and edge-spacing networks. Figure 4.20 illustrates an identical cloud filed with different derived networks. The weight of the incidents links to the larger clouds is remarkably larger in the edge-spacing scheme compared to the center-spacing scheme. Nevertheless, Figure 4.21 demonstrates that the network metrics of these two different schemes are mostly correlated. The metrics that have correlations lower than 0.7 are degree standard deviation ($R = 0.38$), clustering coefficient mean ($R = 0.5$), and maximum betweenness centrality ($R = 0.19$). Compared to Figure 4.13, Figure 4.22 and Figure 4.23 indicate that the degree's mean and standard deviation of the edge-spacing network scheme are considerably more correlated with the first principal component of Martin's metrics. Besides, the metrics algebraic connectivity, degree of assortativity, betweenness centrality mean, and the average shortest path are similarly correlated with the principal components of Martin's metrics in both center and edge-spacing network schemes.

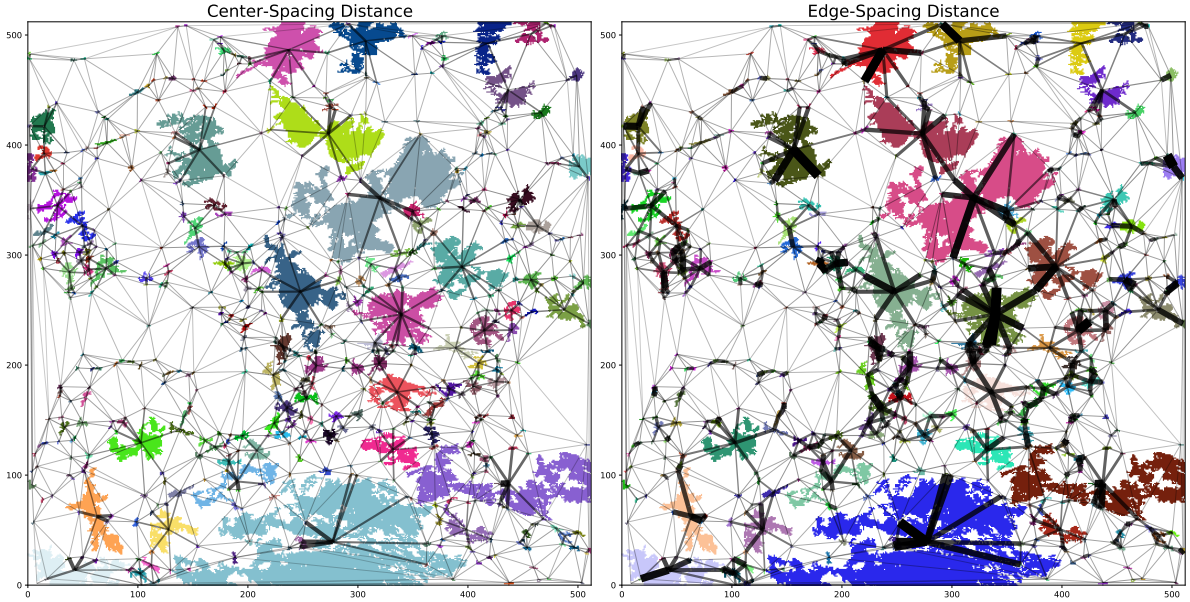


Figure 4.20: Left: The center-spacing network, Right: the edge-spacing network

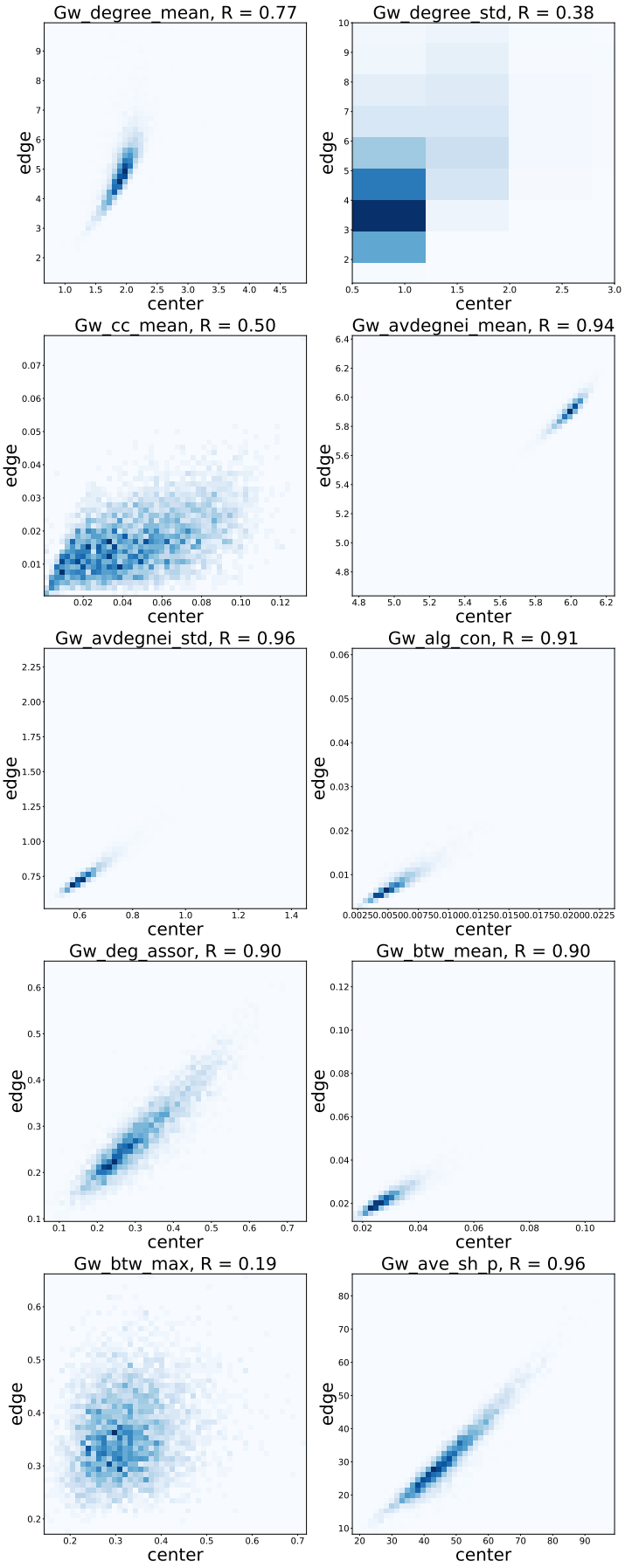


Figure 4.21: The two-dimensional histograms for the important network metrics. x-axis: center-spacing scheme, y-axis: edge-spacing scheme

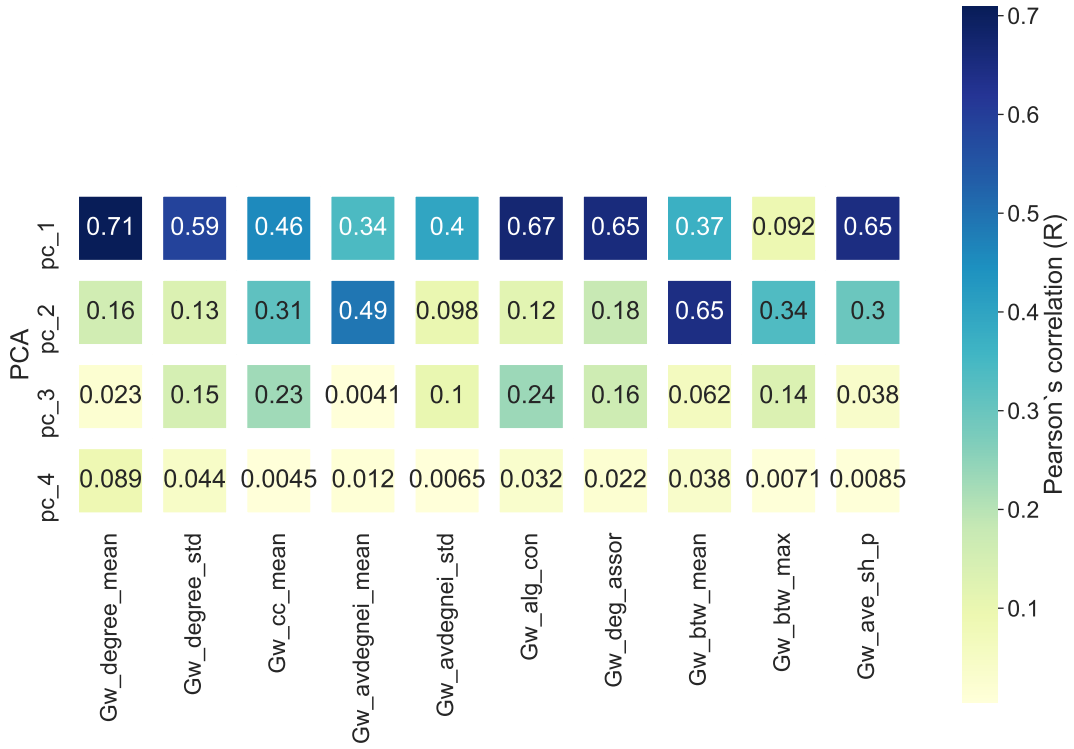


Figure 4.22: Correlation heat-map of weighted network metrics (edge-spacing scheme) with the first four principal components of Martin's metrics

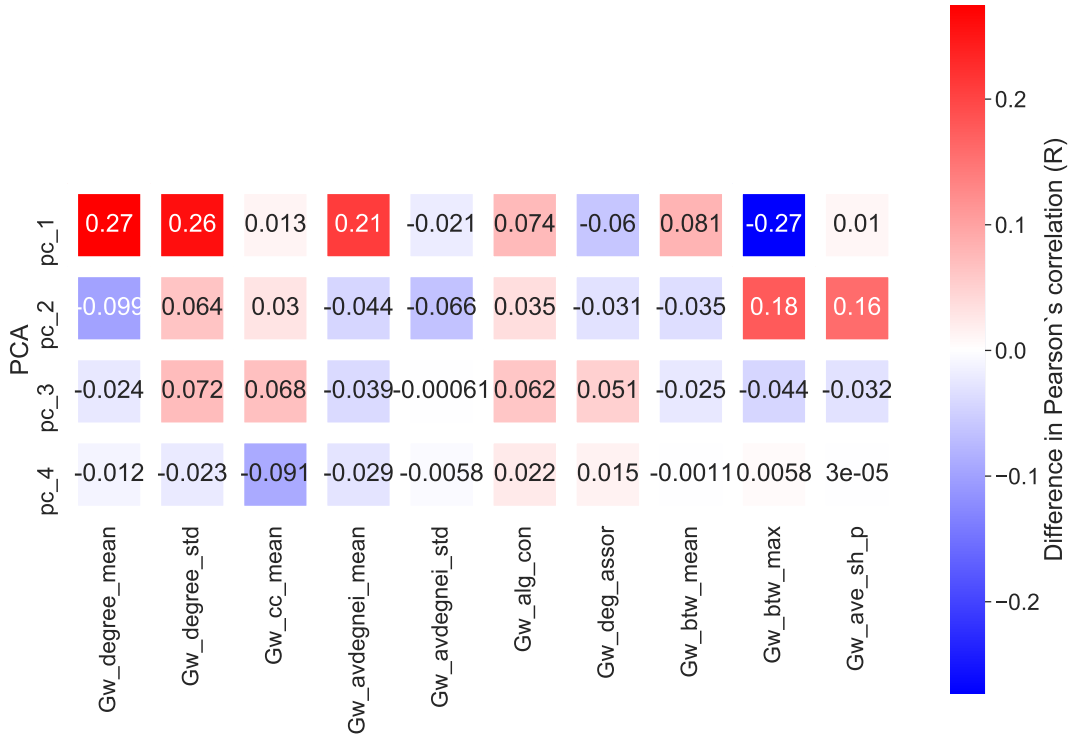


Figure 4.23: The element-wise difference between two correlation heat-maps Figure 4.22 and Figure 4.13

4.5.2. Sensitivity to the Nodes on Boundaries

In this study, when a network is derived from a cloud field, it is inevitably assumed that the clouds located on the boundaries of the field do not have any relationship with the clouds located out of the image provided by the remotely sensed instrument. To investigate how this assumption affects the results of this research, we re-performed the analysis by removing the boundary nodes from the statistical analysis. In other words, the derived networks are the same as the original analysis, but the network metrics are calculated excluding the nodes located on boundaries. In this regard, Figure 4.25 illustrates that the network metrics are not sensitive to either including or excluding the boundary nodes. Because of the high correlation between the metrics with and without boundary nodes the correlation analysis of this sensitivity analysis is presented under section A.3.

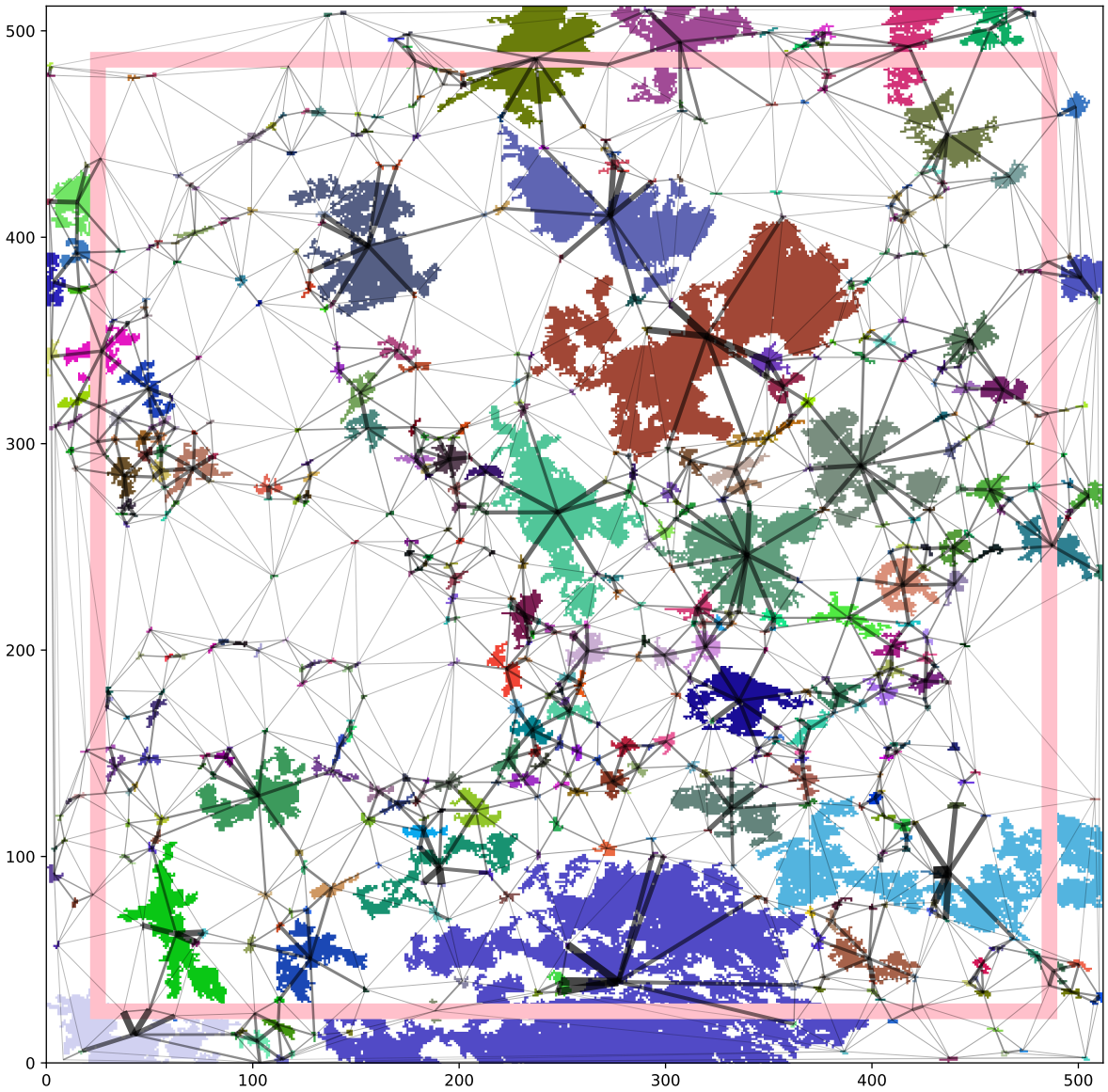


Figure 4.24: Painted cloud field with derived center-spacing network. The pink square encloses the area without the boundary nodes. The size of square is 90 percent of the cloud field domain size.

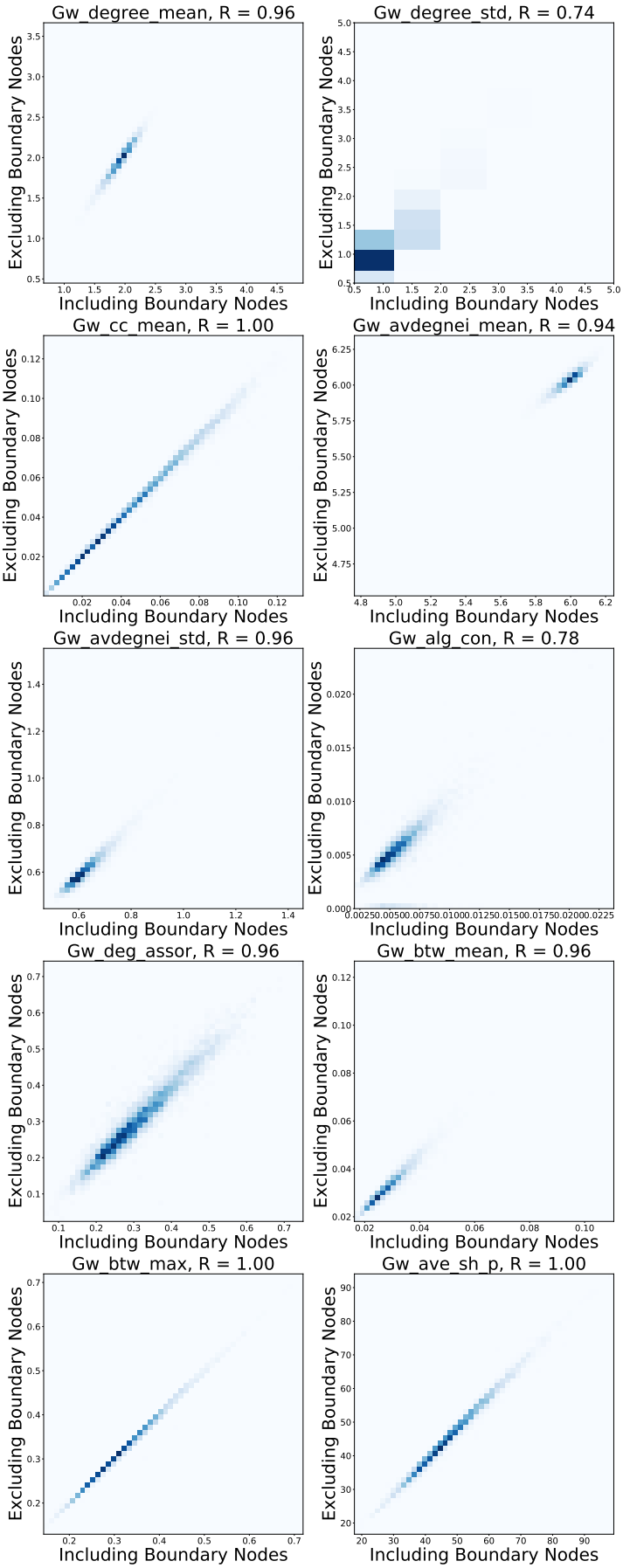


Figure 4.25: The two-dimensional histograms for the important network metrics. x-axis: network metric with the nodes located on boundaries, y-axis: network metric without the nodes located on boundaries

4.6. Discussion Points

In this study, different network schemes are proposed to characterize the interaction between the cloud objects that are present in cumulus cloud fields. As it is demonstrated, how the network is defined remarkably affects all of the associated results. Accordingly, in the weighted network schemes, the weight function is the most indispensable part of the analysis, and therefore the more logical the weight function the more reliable the results. In this regard, the author wants to imply some important points regarding this issue:

- **Process-related Links**
  Active clouds are associated with strong upward motions related to notable physical and thermodynamic processes occurring within the boundary layer. In contrast, the passive clouds are not directly associated with the upward motions, and they may not actively interact with their nearest neighbor. This led the author to mention that defining the weight function based on the physical and thermodynamic processes coupled with clouds may lead to a network that more faithfully characterize the interaction between cumulus cloud objects.

- **Satellite or Large-Eddy Simulation Data**
  To define a process-related weight function, satellite images may not be useful, because it might not be possible to deduce the wind speed and related updraft motions from the satellite data. However, with the help of detailed cloud models such as Dutch Atmospheric Large-Eddy Simulation (DALES) Heus et al. (2010) model, it is possible to obtain the characteristics of the updraft motions and the other related processes.

- **Temporal Analysis of Networks**
  Employing the models like DALES can be fruitful to investigate the evolution of clouds organization and its relationship with the associated meteorological conditions Janssens et al. (2021). Analyzing the derived networks in time would be tremendously worthwhile in capturing the dynamics of the networks and that of cloud fields. Therefore, the next step of this research would be devoted to analyzing the cloud fields in not only space but also time. The simplest idea to start the analysis of derived networks from the cloud fields in time is to plot the time series of network metrics to see how a network metric changes in a diurnal cycle. Besides, auto-correlation analysis of LES data would help us to check the randomness of a network metric.

5

Conclusion

Clouds, in particular, shallow clouds over subtropical oceans are the most prominent source of uncertainty in climate projections (Schneider et al., 2017). Scientists in atmospheric and climate sciences have not yet concluded whether shallow clouds contribute to a lower ECS (cooling effect) or a higher ECS (warming effect) (Schneider et al., 2017;Nuijens and Siebesma, 2019). Besides, shallow cumulus clouds show diverse spatial patterns over subtropical oceans, and the occurrence probability of these patterns can impact the global radiation budget of the Earth and consequently the climate feedback of subtropical marine clouds (Janssens et al., 2021).

In this regard, numerous metrics have been proposed to quantitatively characterize the patterns and organizations of clouds. Most of the proposed organization metrics are the bulk parameters of the cloud fields. Correspondingly, the motivation of this research is to find a metric that can characterize the mutual arrangement between cloud objects. To this end, network theory as a worthwhile tool is employed for proposing metrics to quantify how cloud objects interact with and coordinate concerning each other. We derive weighted networks from 5000 cloud fields from the NASA worldview website from 2002 to 2020 within December to May. For each cloud field, different network metrics are calculated. To understand how distinctly the proposed network metrics imply for cloud organization, we compare them with the previously defined organization metrics that were collected by Janssens et al. (2021).

Compared to the previously defined metrics, most of the network metrics are correlated with the non-network metrics. The reason for the large values of correlation between the network and non-network metrics may be the fact that our weighed network scheme is defined based on the area of and the distance between the cloud objects. Among the tightly correlated network metrics with the existing metrics, two pairs of metrics have correlation values larger than 0.7 including mean betweenness centrality and fractal dimension ($R = -0.73$) and degree of assortativity and size exponent ($R = -0.72$). Besides, it is found that none of the network metrics can replace one of the first four principal components of the existing metrics. Conversely, the linear combination of the principal components of previously defined metrics cannot adequately capture the variation in two network metrics, mean of clustering coefficient and degree standard deviation. Thus, the aforementioned metrics can uniquely capture variations that have not been captured by previously defined metrics in cumulus cloud organization. Accordingly, a large value of the average clustering coefficient is associated with the presence of a large number of triangles with high-weighted sides (links) in the derived networks from cloud scenes. This is associated with the presence of either large clouds located on corners of triangles or relatively small clouds closely coagulated. Besides, degree standard deviation simultaneously measures the homogeneity of the cloud size distribution and the distance between nearest neighboring clouds distribution. Additionally, the first sensitivity analysis reveals that our main results are not sensitive if the distance term is either center-spacing or edge-spacing. Similarly, the second sensitivity analysis implies that our network metrics are not susceptible to removing the nodes located on the boundaries of the cloud field from the statistical analyses.

Moreover, one should note that the most vital part of this research is how to define the network and related weight function. In consequence, this issue notably controls all of the outputs of this study. In this research, the weighted networks are defined based on only geometric properties of the cloud field, clouds' area, and the distance between the nearest neighbors. Therefore, it could be a difficult task to deduce direct physical process-related conclusions from the geometric-based results. Hence, this study can be improved by utilizing the network theory to define the process-related weight function. On the other hand, instead of using satellite snapshots, the LES results are comparably fruitful to distinguish active clouds from the passive clouds in the cloud scene. Eventually, the future work of this research will be devoted to analyzing the dynamics of the networks in time to encapsulate the evolution of clouds and their interaction not only in the spatial domain but also in the temporal domain. To this end, one can start the temporal analysis with auto-correlation analysis of LES data to check the randomness of the network metrics.

Bibliography

Allen-Perkins, A., Pastor, J. M., and Estrada, E. (2017). Two-walks degree assortativity in graphs and networks. *Applied Mathematics and Computation*, 311:262–271.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101(11):3747–3752.

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1-3):1–101.

Bony, S. and Dufresne, J.-L. (2005). Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophysical Research Letters*, 32(20).

Bony, S., Schulz, H., Vial, J., and Stevens, B. (2020). Sugar, gravel, fish, and flowers: Dependence of mesoscale patterns of trade-wind clouds on environmental conditions. *Geophysical research letters*, 47(7):e2019GL085988.

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145.

Cahalan, R. F. and Joseph, J. H. (1989). Fractal statistics of cloud fields. *Monthly weather review*, 117(2):261–272.

Cess, R. D., Potter, G., Blanchet, J., Boer, G., Del Genio, A., Deque, M., Dymnikov, V., Galin, V., Gates, W., Ghan, S., et al. (1990). Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *Journal of Geophysical Research: Atmospheres*, 95(D10):16601–16615.

Ding, B., Li, C., Zhang, M., Lu, G., and Ji, F. (2014). Numerical analysis of percolation cluster size distribution in two-dimensional and three-dimensional lattices. *The European Physical Journal B*, 87(8):1–8.

Glassmeier, F. and Feingold, G. (2017). Network approach to patterns in stratocumulus clouds. *Proceedings of the National Academy of Sciences*, 114(40):10578–10583.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Heus, T., van Heerwaarden, C. C., Jonker, H. J. J., Pier Siebesma, A., Axelsen, S., van den Dries, K., Geoffroy, O., Moene, A. F., Pino, D., de Roode, S. R., and Vilà-Guerau de Arellano, J. (2010). Formulation of the dutch atmospheric large-eddy simulation (dales) and overview of its applications. *Geoscientific Model Development*, 3(2):415–444.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Janssens, M., Vilà-Guerau de Arellano, J., Scheffer, M., Antonissen, C., Siebesma, A. P., and Glassmeier, F. (2021). Cloud patterns in the trades have four interpretable dimensions. *Geophysical Research Letters*, page e2020GL091001.

Laar, T. W. v. (2019). *Spatial patterns in shallow cumulus cloud populations over a heterogeneous surface*. PhD thesis, Universität zu Köln.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2003a). Mixing patterns in networks. *Physical review E*, 67(2):026126.

Newman, M. E. (2003b). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Nuijens, L., Serikov, I., Hirsch, L., Lonitz, K., and Stevens, B. (2014). The distribution and variability of low-level cloud in the north atlantic trades. *Quarterly Journal of the Royal Meteorological Society*, 140(684):2364–2374.

Nuijens, L. and Siebesma, A. P. (2019). Boundary layer clouds and convection over subtropical oceans in our current and in a warmer climate. *Current Climate Change Reports*, 5(2):80–94.

Onnela, J.-P., Saramäki, J., Kertész, J., and Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., and Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, 7(1):3–5.

Schulz, H., Eastman, R., and Stevens, B. (2021). Characterization and evolution of organized shallow convection in the trades. *Journal of Geophysical Research: Atmospheres*.

Stevens, B., Bony, S., Brogniez, H., Hentgen, L., Hohenegger, C., Kiemle, C., L'Ecuyer, T. S., Naumann, A. K., Schulz, H., Siebesma, P. A., et al. (2020). Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds. *Quarterly Journal of the Royal Meteorological Society*, 146(726):141–152.

Tipping, M. E. and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482.

Van Mieghem, P. (2014). *Performance analysis of complex networks and systems*. Cambridge University Press.

Vial, J., Dufresne, J.-L., and Bony, S. (2013). On the interpretation of inter-model spread in cmip5 climate sensitivity estimates. *Climate Dynamics*, 41(11-12):3339–3362.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wallace, J. and Hobbs, P. (2006). *Atmospheric science: an introductory survey*, volume 92. Elsevier.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

White, B., Buchanan, A., Birch, C., Stier, P., and Pearson, K. (2018). Quantifying the effects of horizontal grid length and parameterized convection on the degree of convective organization using a metric of the potential for convective interaction. *Journal of the Atmospheric Sciences*, 75(2):425–450.

A

Appendices

A.1. Sort Plots of Network Metrics

A.1.1. Weighted Network Metrics

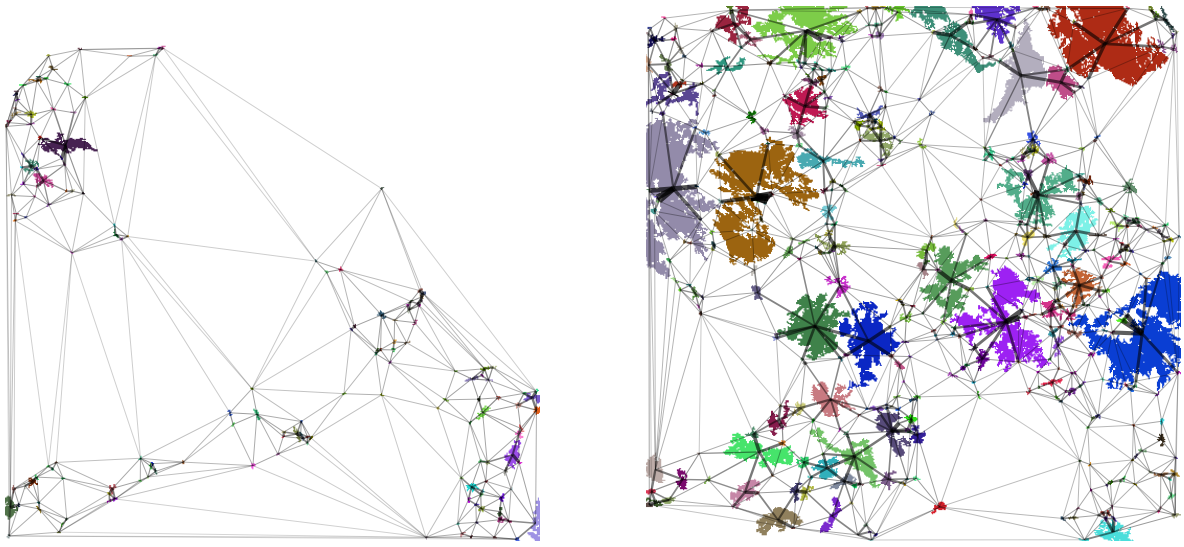


Figure A.1: The cloud scenes are ordered by the degree mean value from left to right.

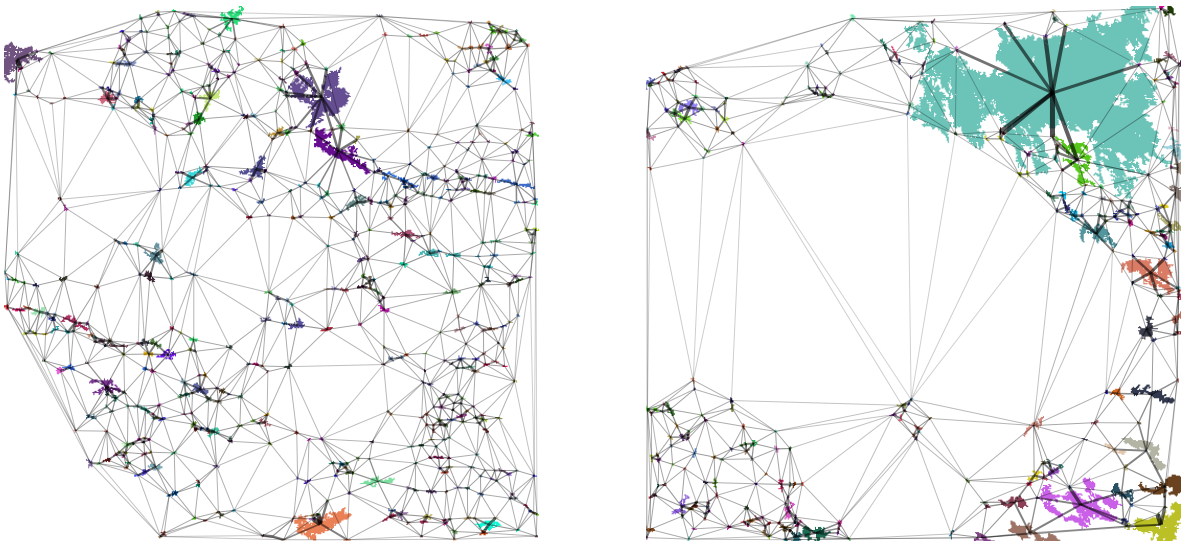


Figure A.2: The cloud scenes are ordered by the degree standard deviation value from left to right.

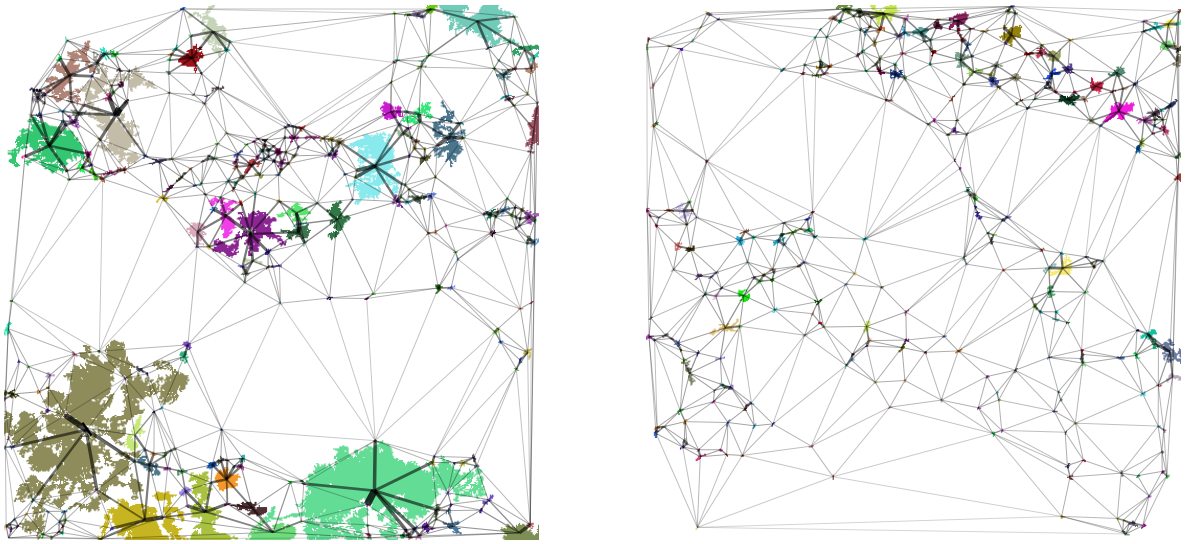


Figure A.3: The cloud scenes are ordered by the clustering coefficient mean value from left to right.

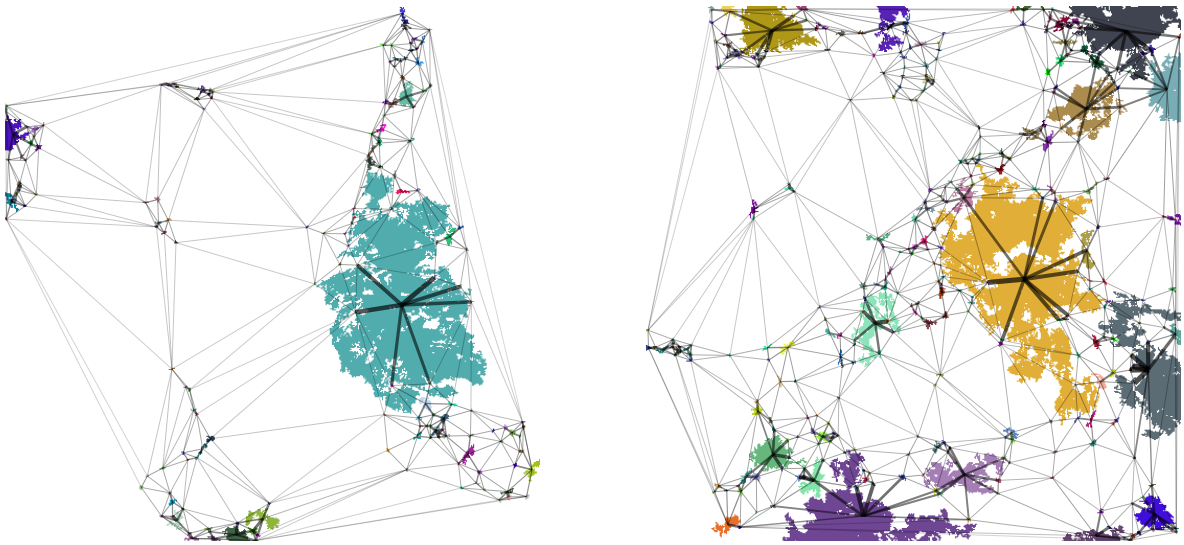


Figure A.4: The cloud scenes are ordered by the average degree of neighbors mean value from left to right.

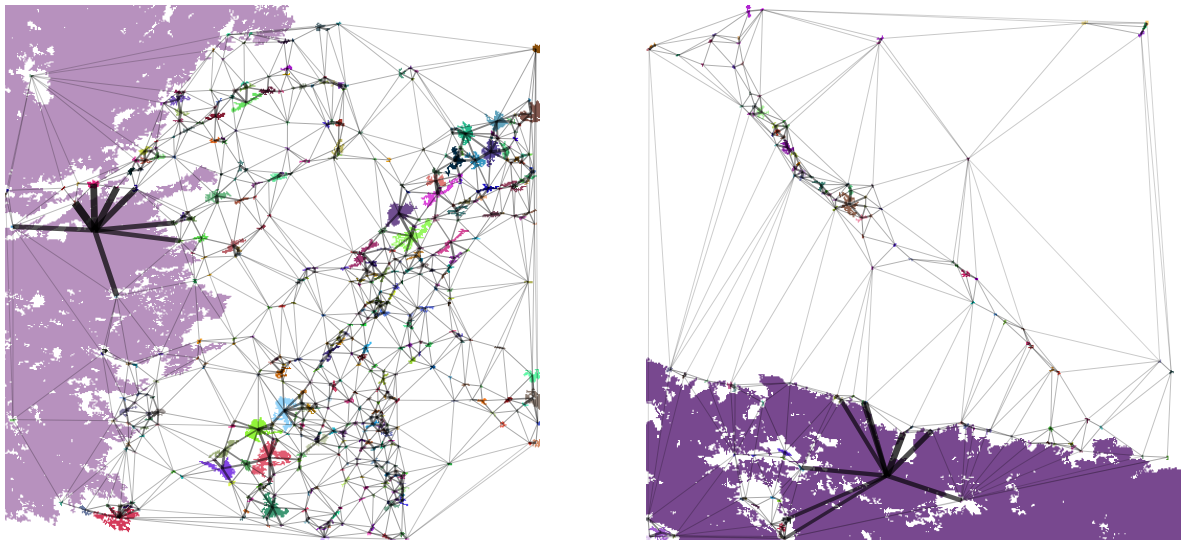


Figure A.5: The cloud scenes are ordered by the average degree of neighbors standard deviation value from left to right.

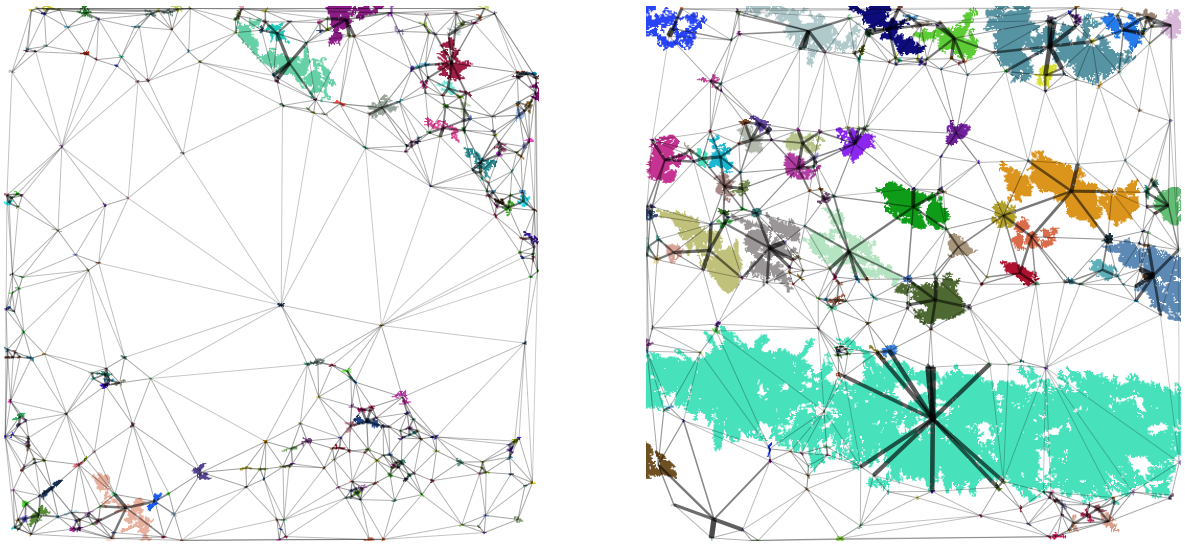


Figure A.6: The cloud scenes are ordered by the algebraic connectivity value from left to right.

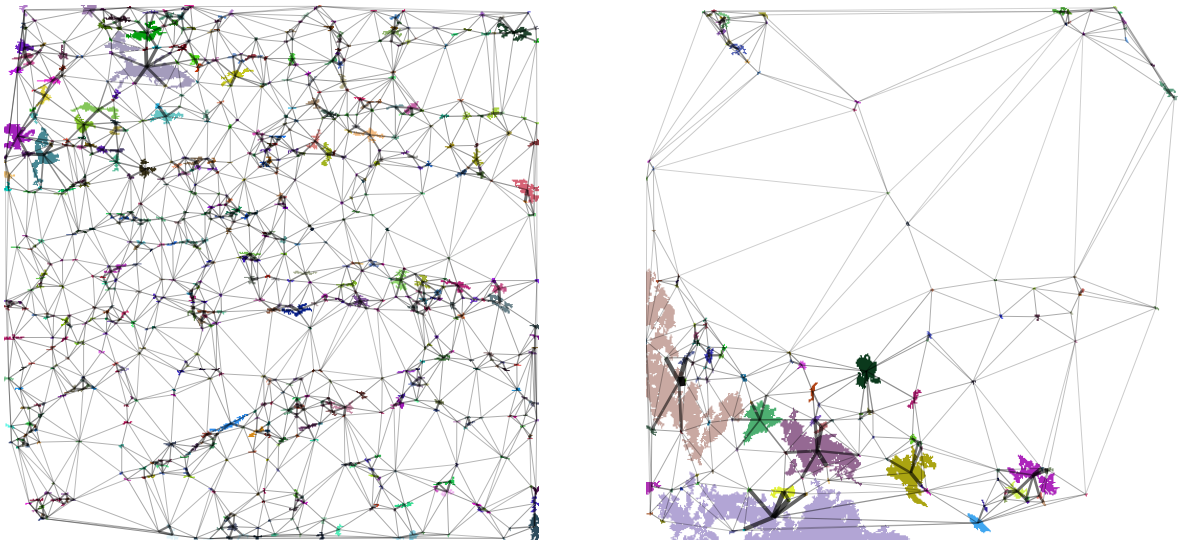


Figure A.7: The cloud scenes are ordered by the betweenness centrality mean value from left to right.

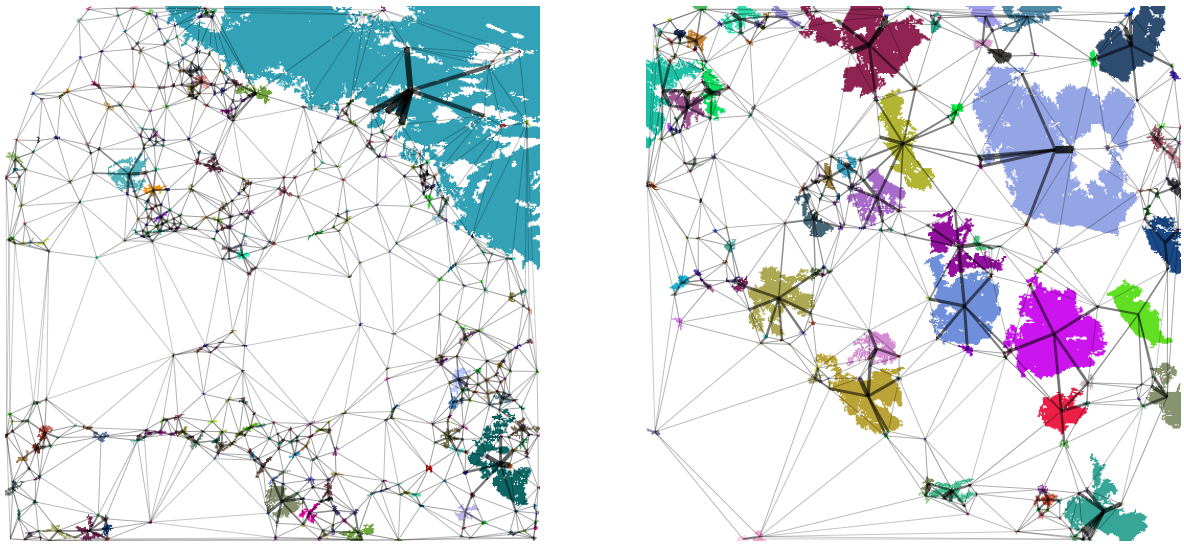


Figure A.8: The cloud scenes are ordered by the betweenness centrality max value from left to right.

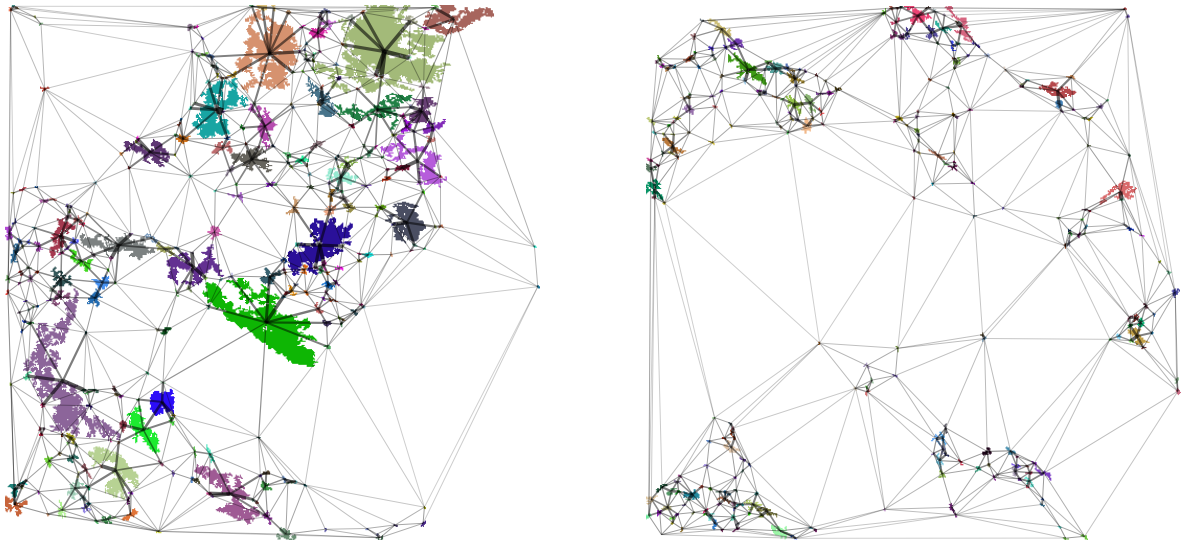


Figure A.9: The cloud scenes are ordered by the average shortest path value from left to right.

A.1.2. Removed-Edge Network Metrics

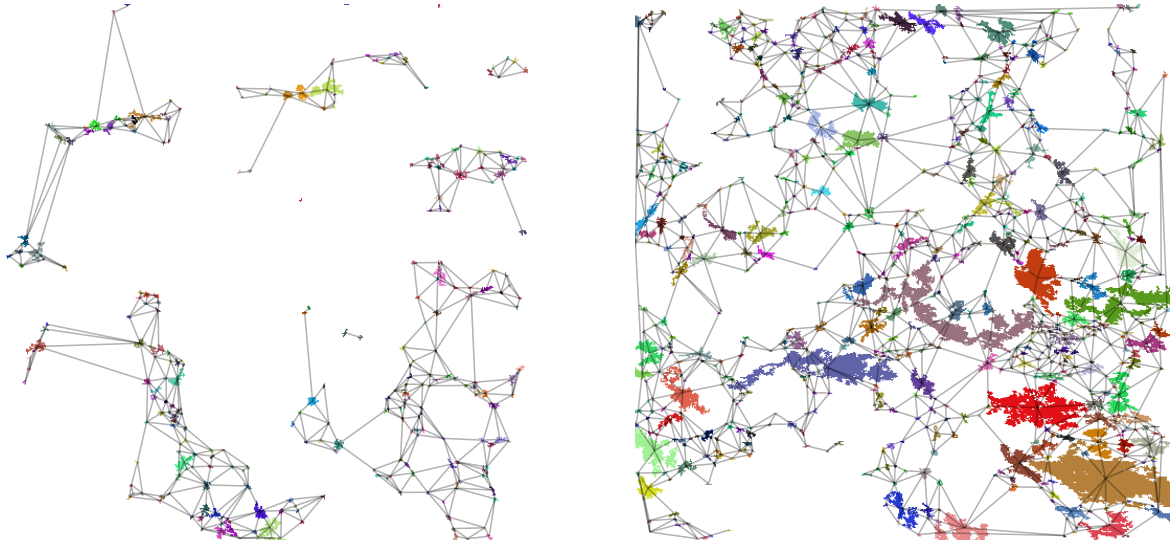


Figure A.10: The cloud scenes are ordered by the degree mean value from left to right.

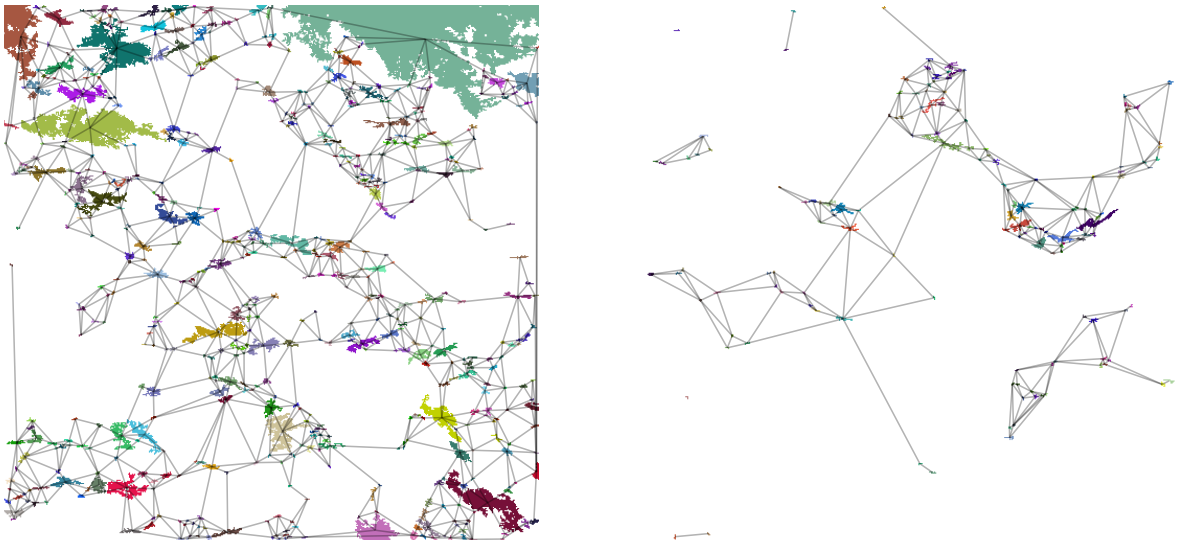


Figure A.11: The cloud scenes are ordered by the degree standard deviation value from left to right.

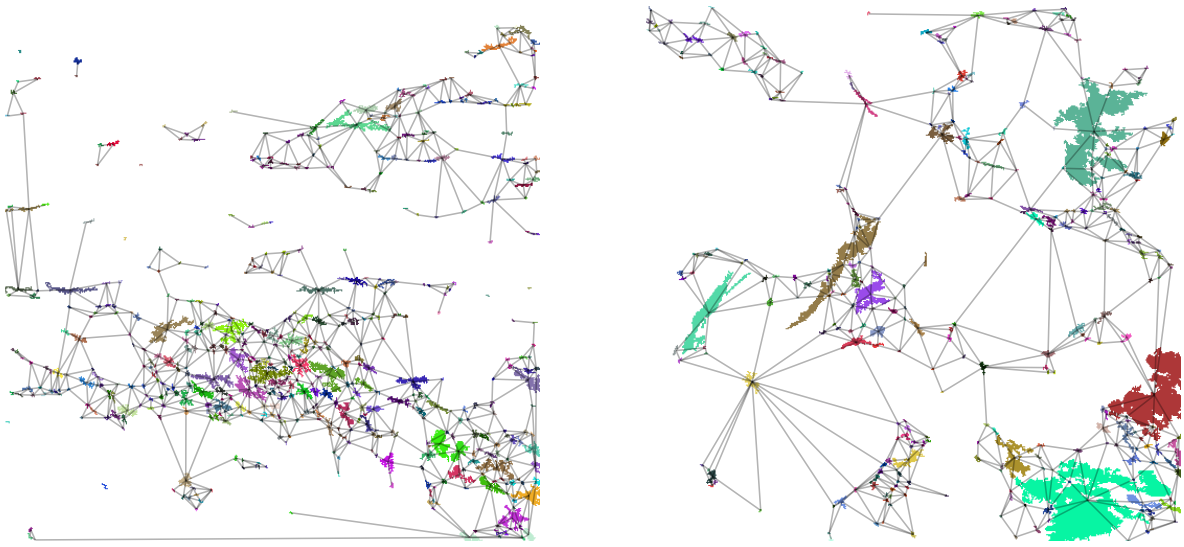


Figure A.12: The cloud scenes are ordered by the degree skewness value from left to right.

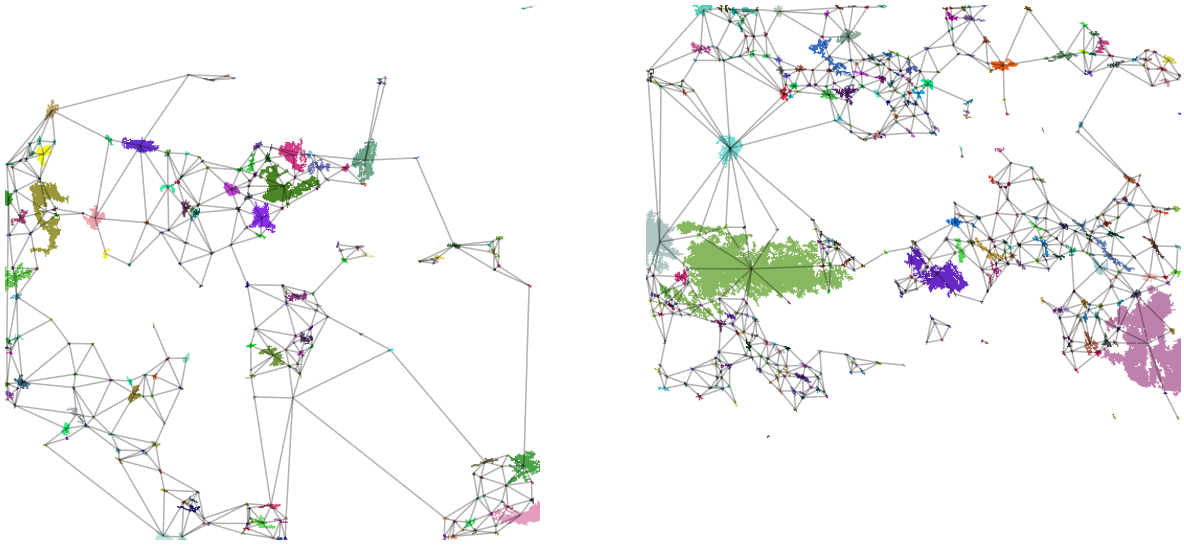


Figure A.13: The cloud scenes are ordered by the maximum degree value from left to right.

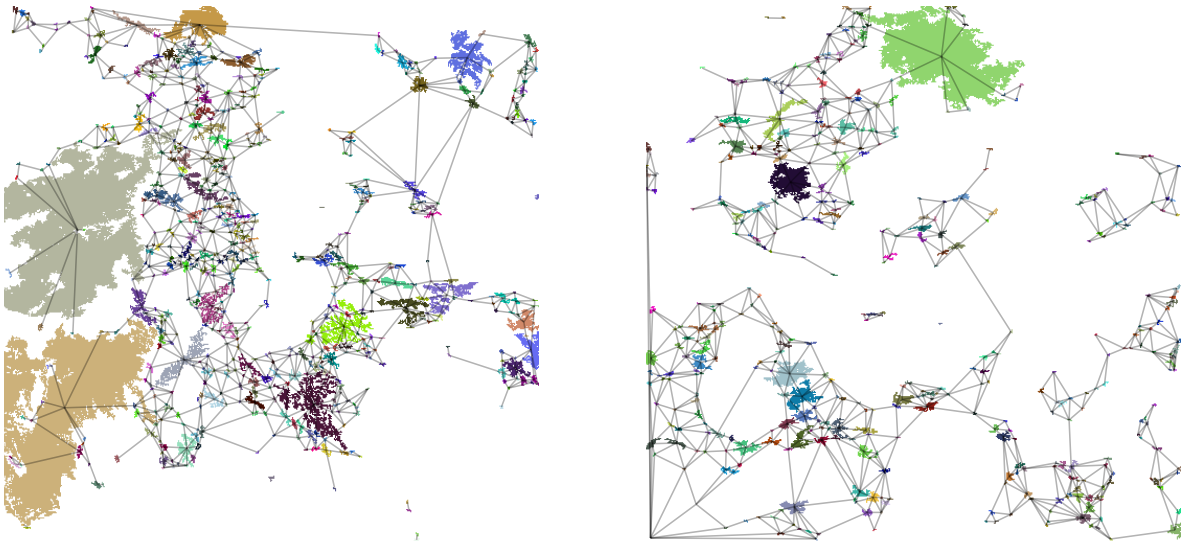


Figure A.14: The cloud scenes are ordered by the clustering coefficient mean value from left to right.

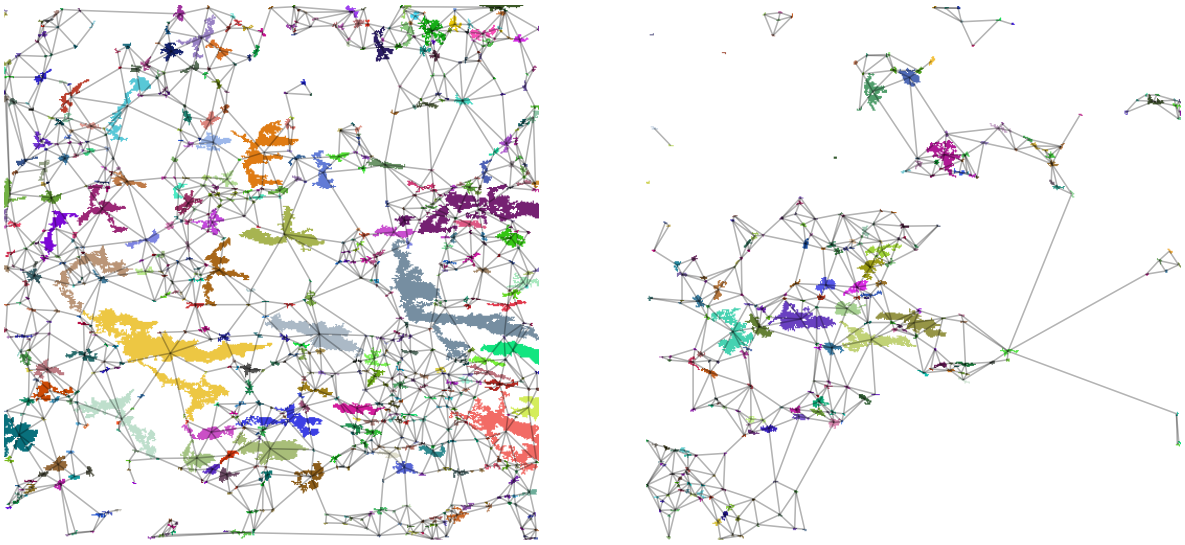


Figure A.15: The cloud scenes are ordered by the clustering coefficient standard deviation value from left to right.

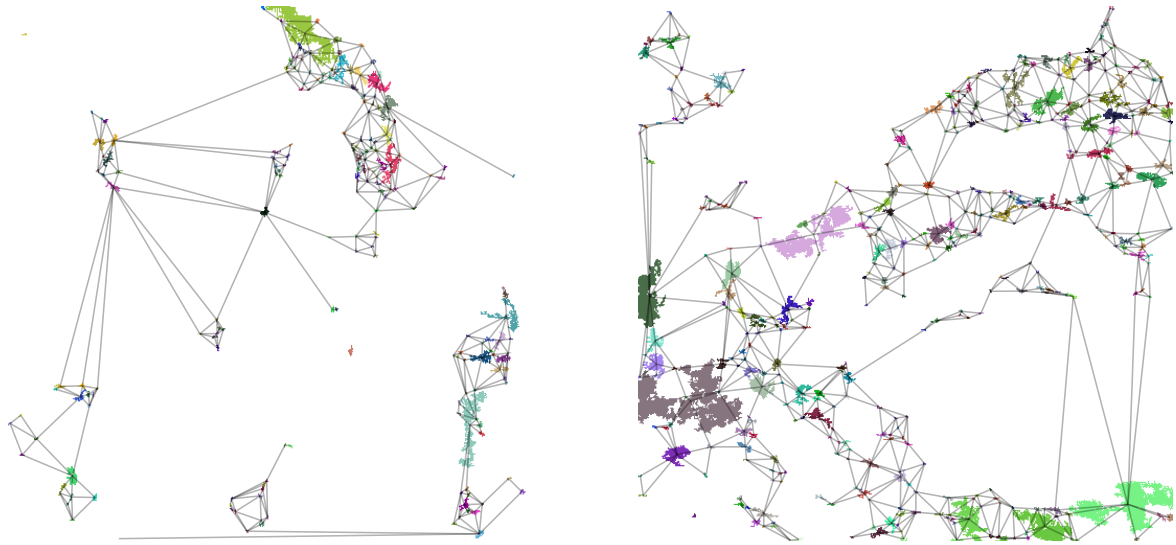


Figure A.16: The cloud scenes are ordered by the clustering coefficient skewness value from left to right.

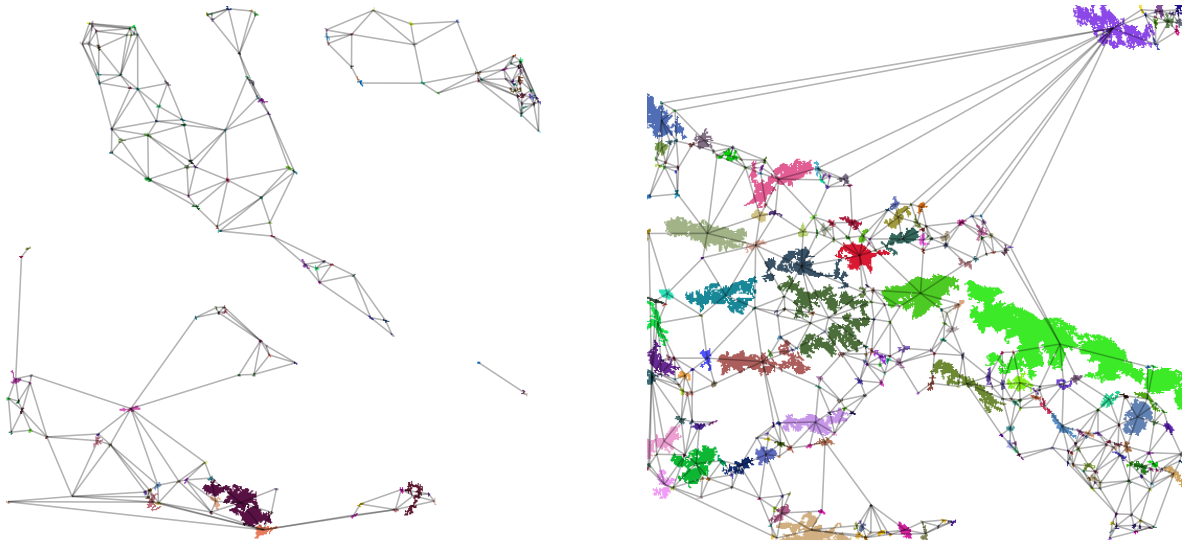


Figure A.17: The cloud scenes are ordered by the average degree of neighbors mean value from left to right.

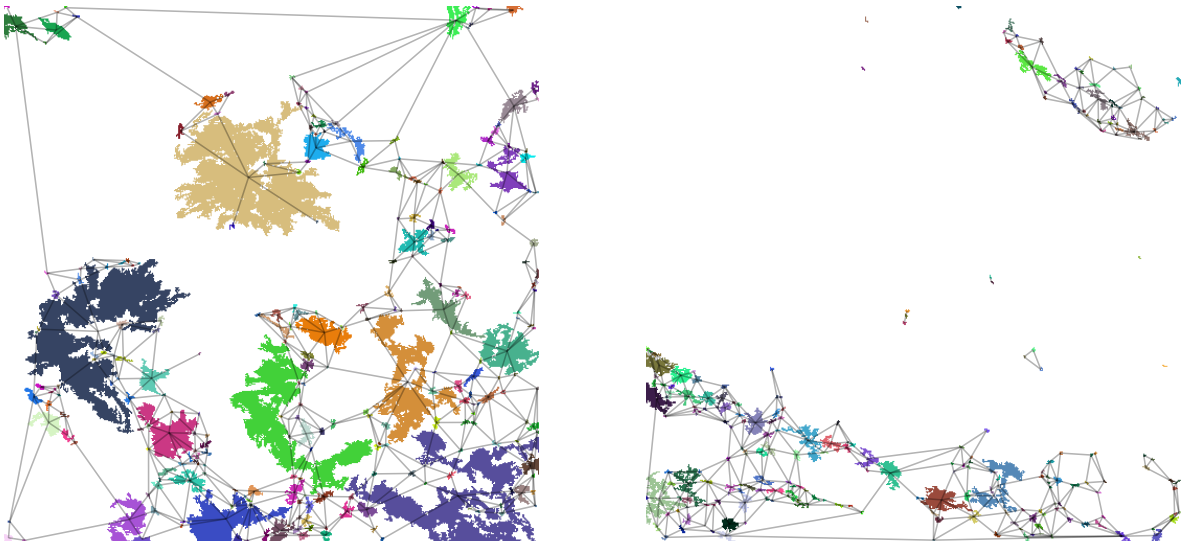


Figure A.18: The cloud scenes are ordered by the average degree of neighbors standard deviation value from left to right.

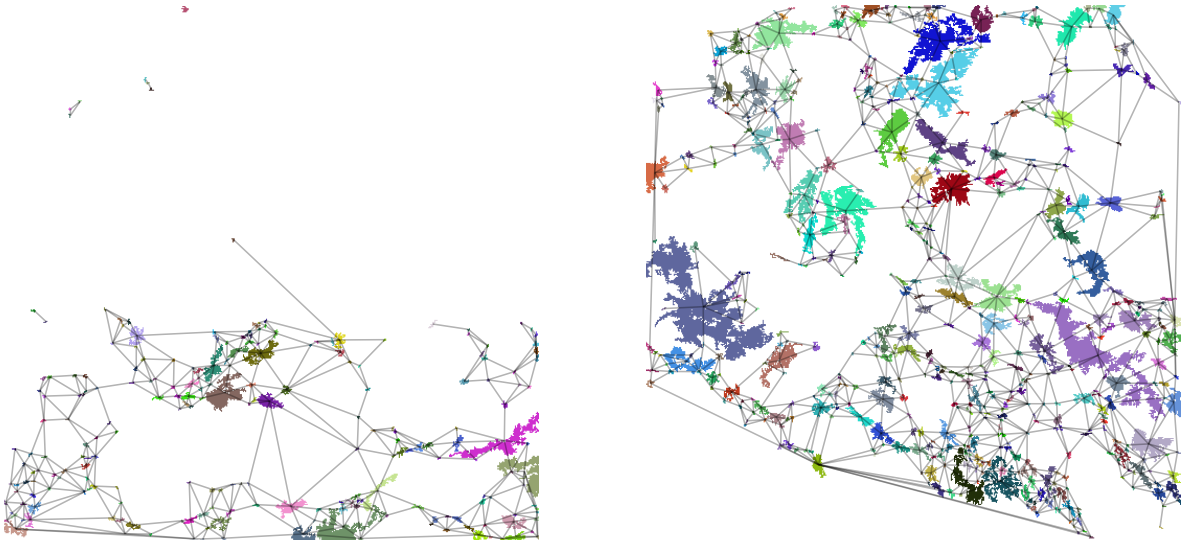


Figure A.19: The cloud scenes are ordered by the average degree of neighbors skewness value from left to right.

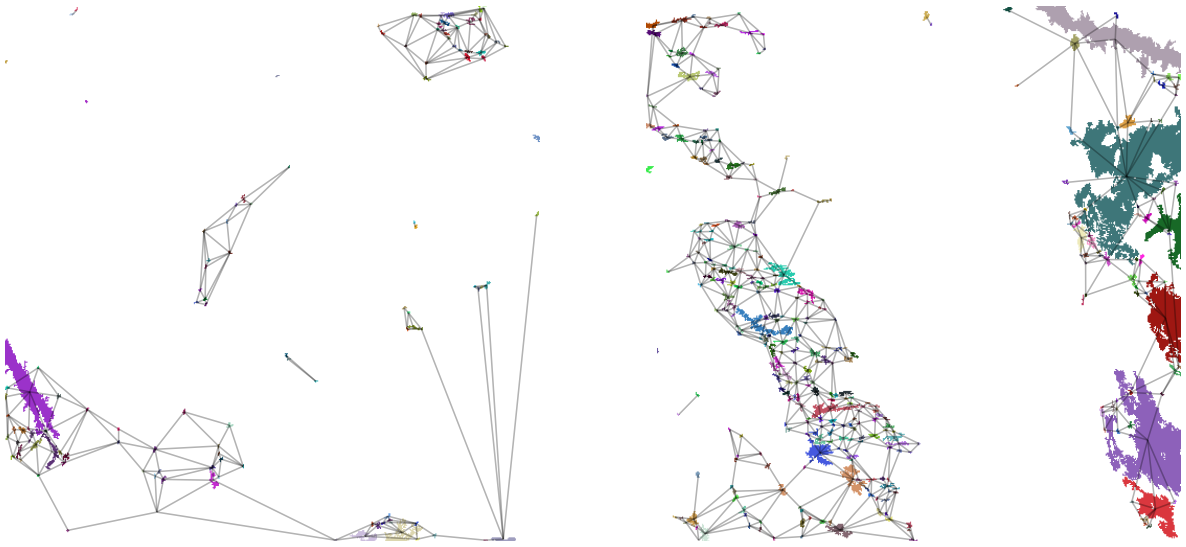


Figure A.20: The cloud scenes are ordered by the maximum average degree of neighbors value from left to right.

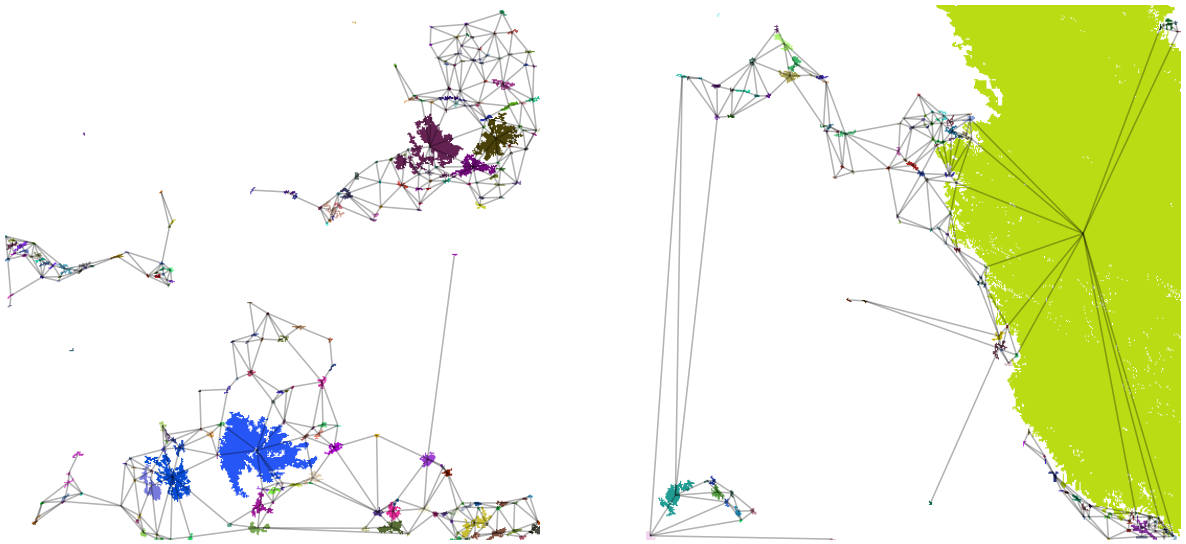


Figure A.21: The cloud scenes are ordered by the betweenness centrality mean value from left to right.

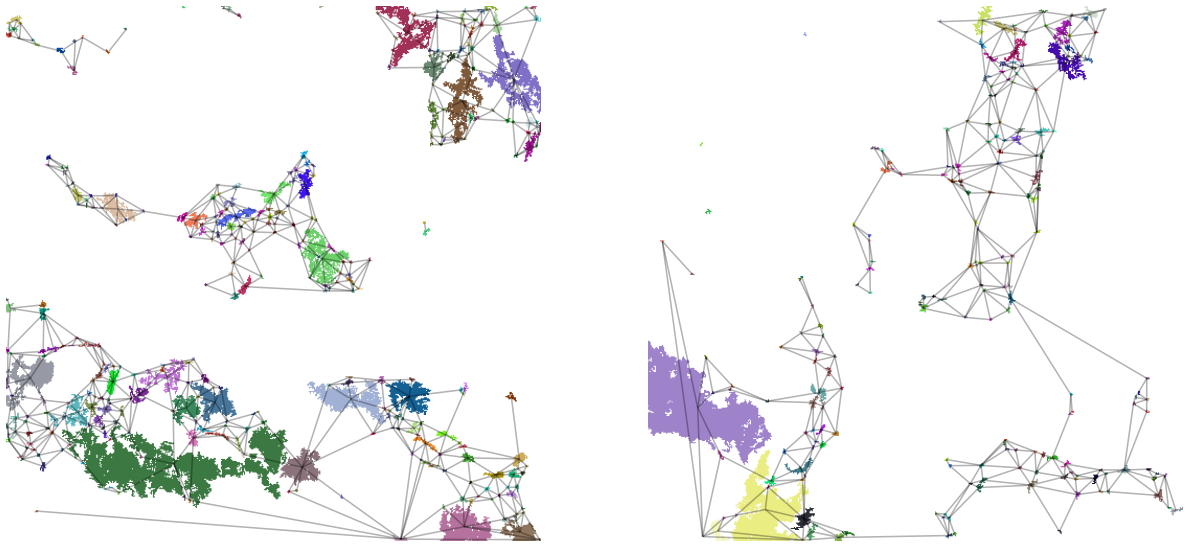


Figure A.22: The cloud scenes are ordered by the betweenness centrality standard deviation value from left to right.

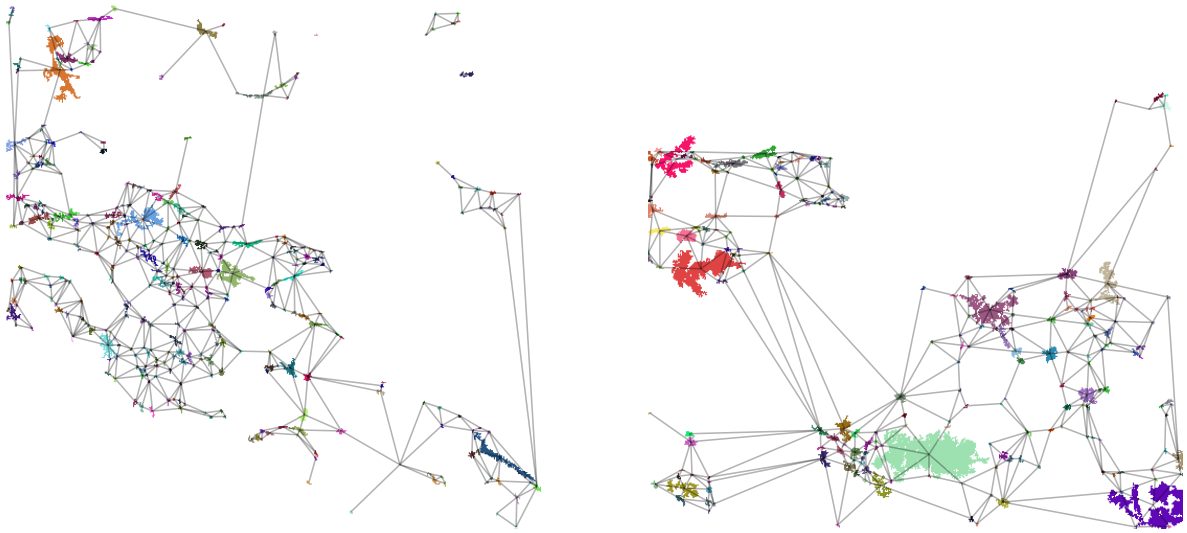


Figure A.23: The cloud scenes are ordered by the betweenness centrality skewness value from left to right.

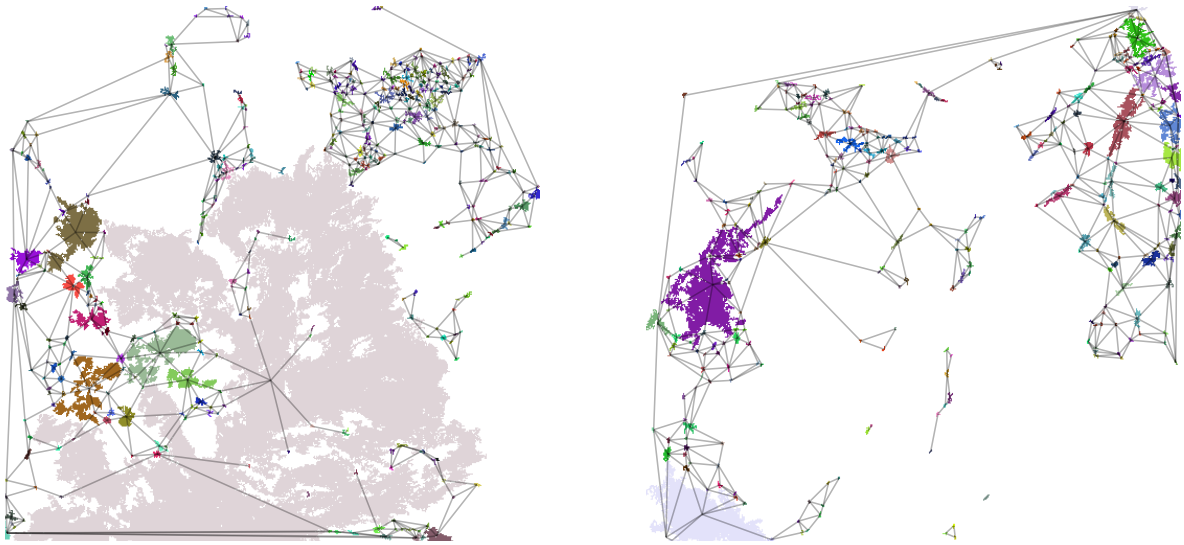


Figure A.24: The cloud scenes are ordered by the maximum betweenness centrality value from left to right.

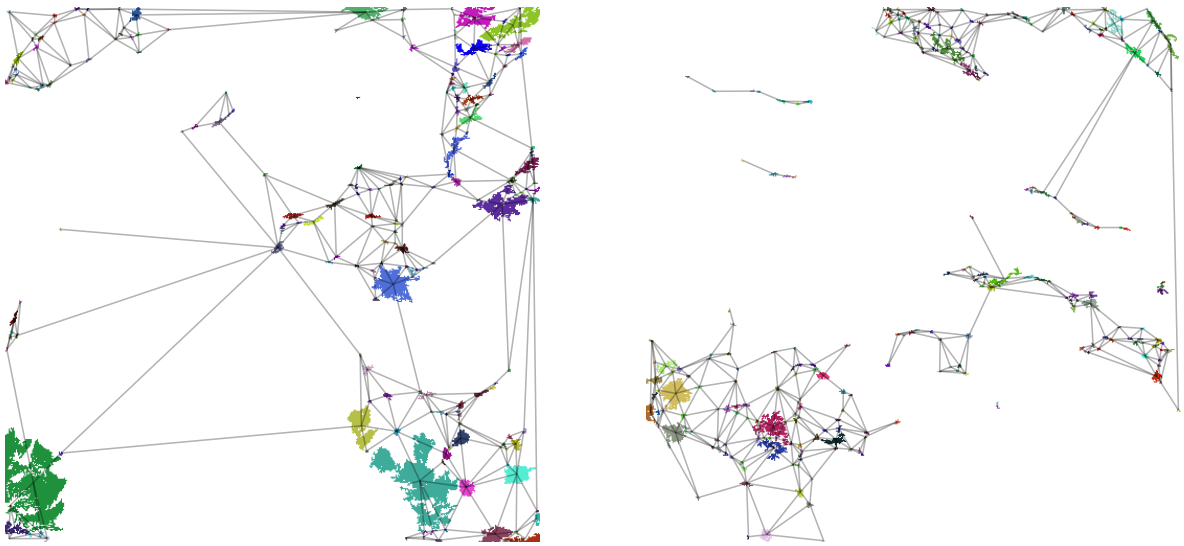


Figure A.25: The cloud scenes are ordered by the degree of assortativity value from left to right.

A.2. Sensitivity to the Distance Term

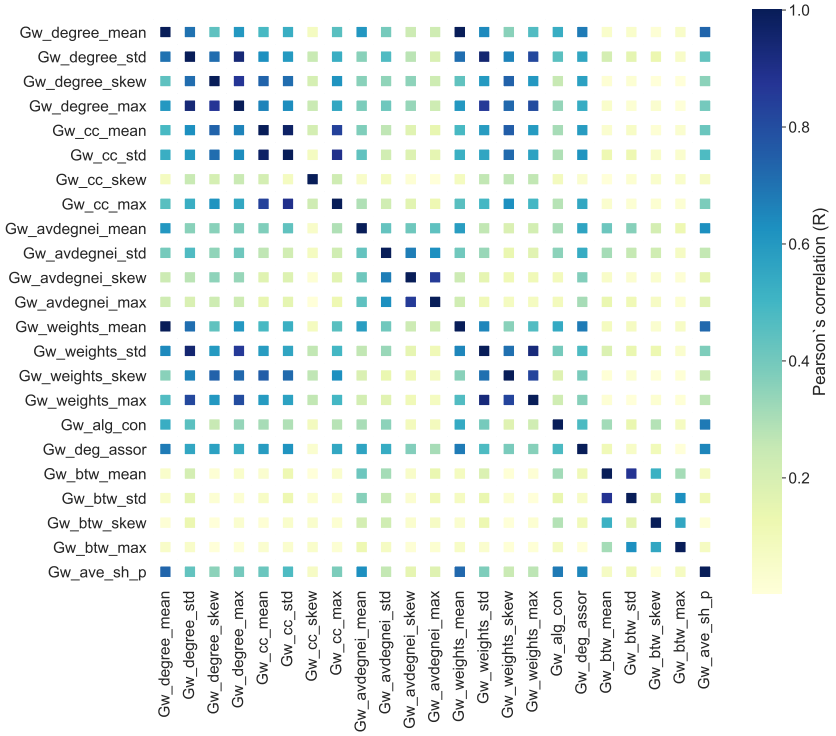


Figure A.26: Correlation heat-map of all edge-spacing network metrics

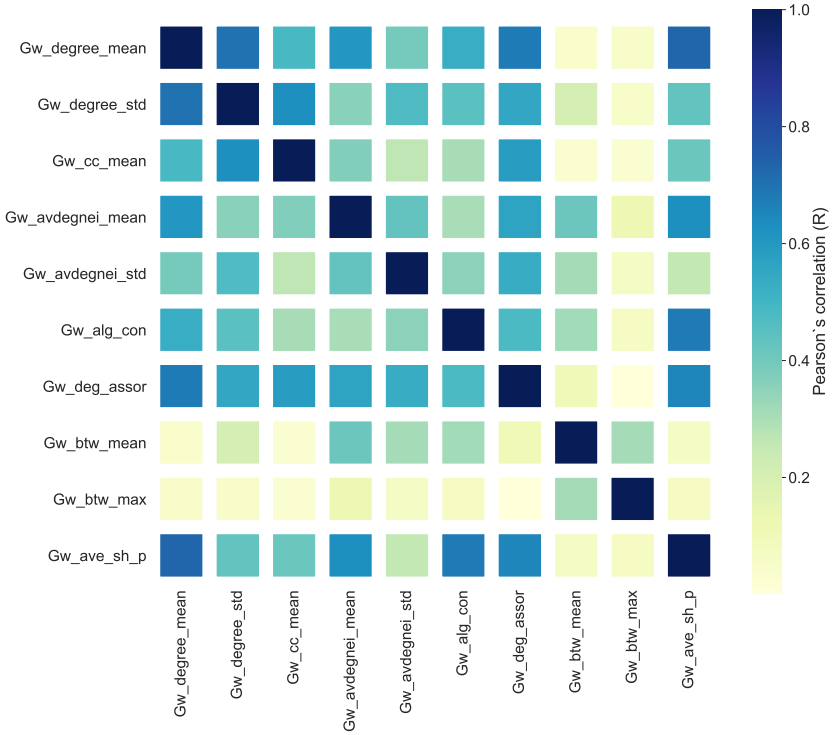


Figure A.27: Correlation heat-map of reduced metrics for the edge-spacing network scheme

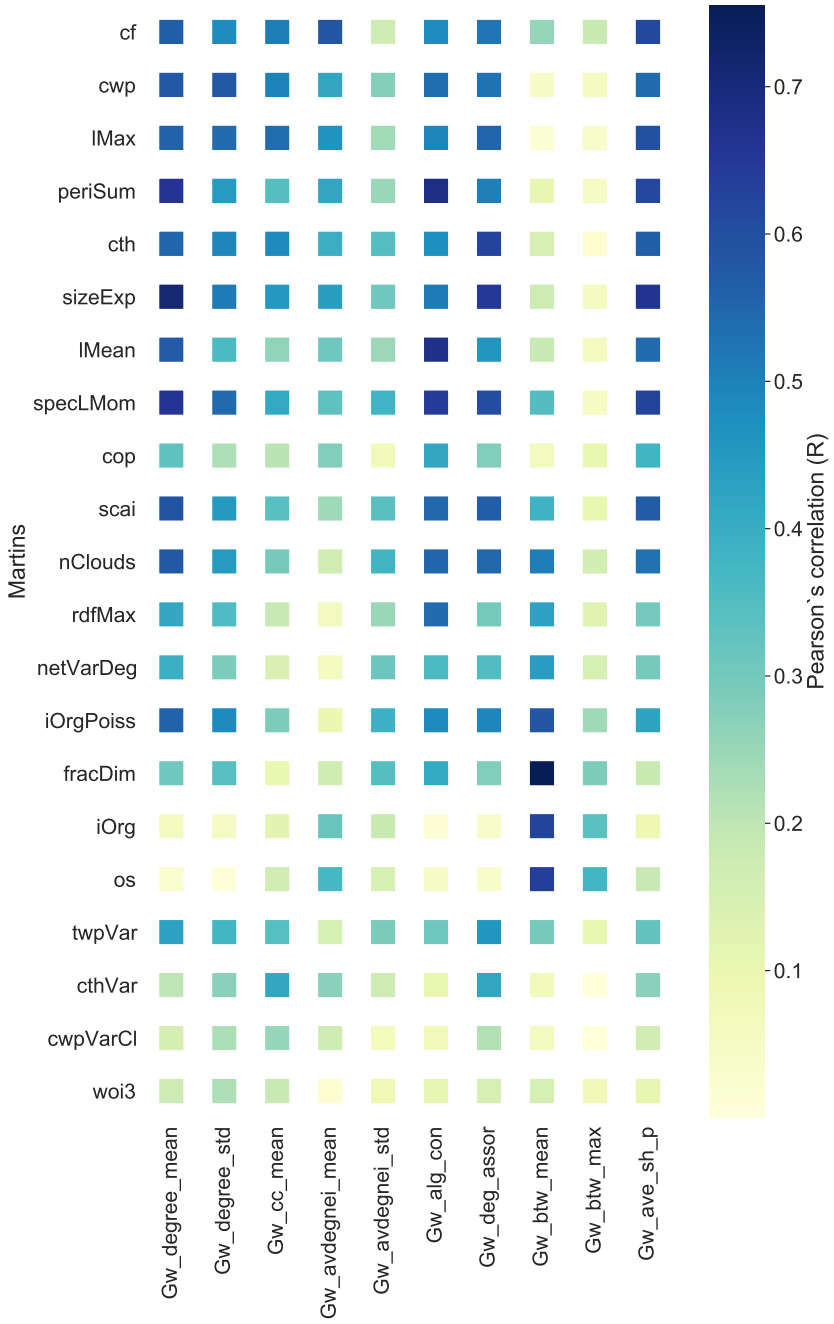


Figure A.28: Correlation heat-map of edge-spacing network metrics with Martin's metrics

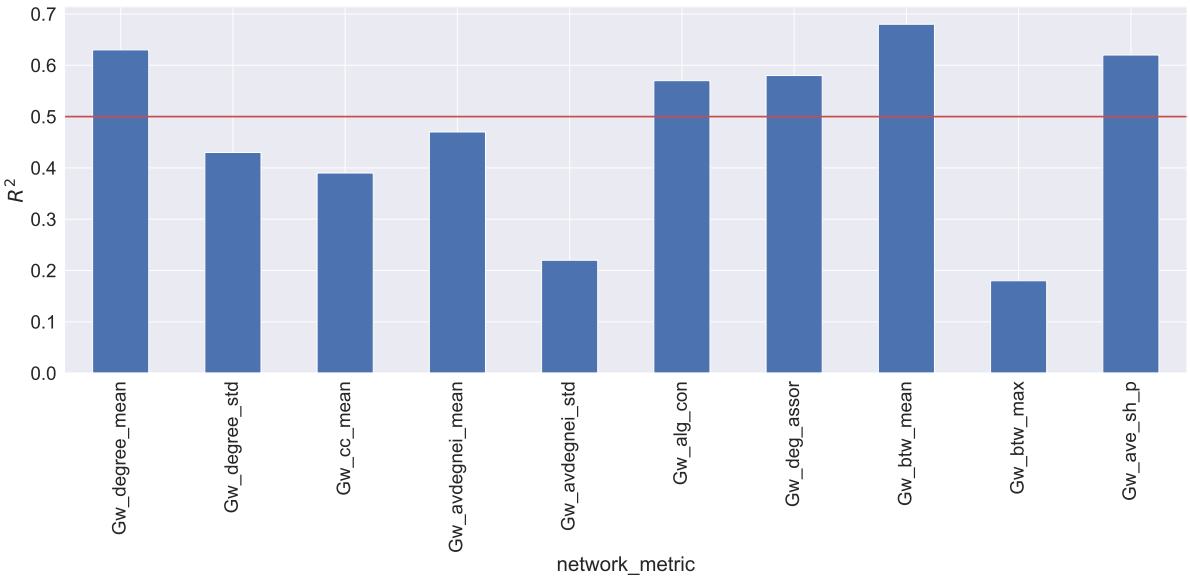


Figure A.29: The figure shows the R-squared values of each multiple linear regression model. In each model, the target feature is one of the edge-spacing network metrics and the regressors are principal components of Martin's metrics.

A.3. Sensitivity to the Nodes on Boundaries

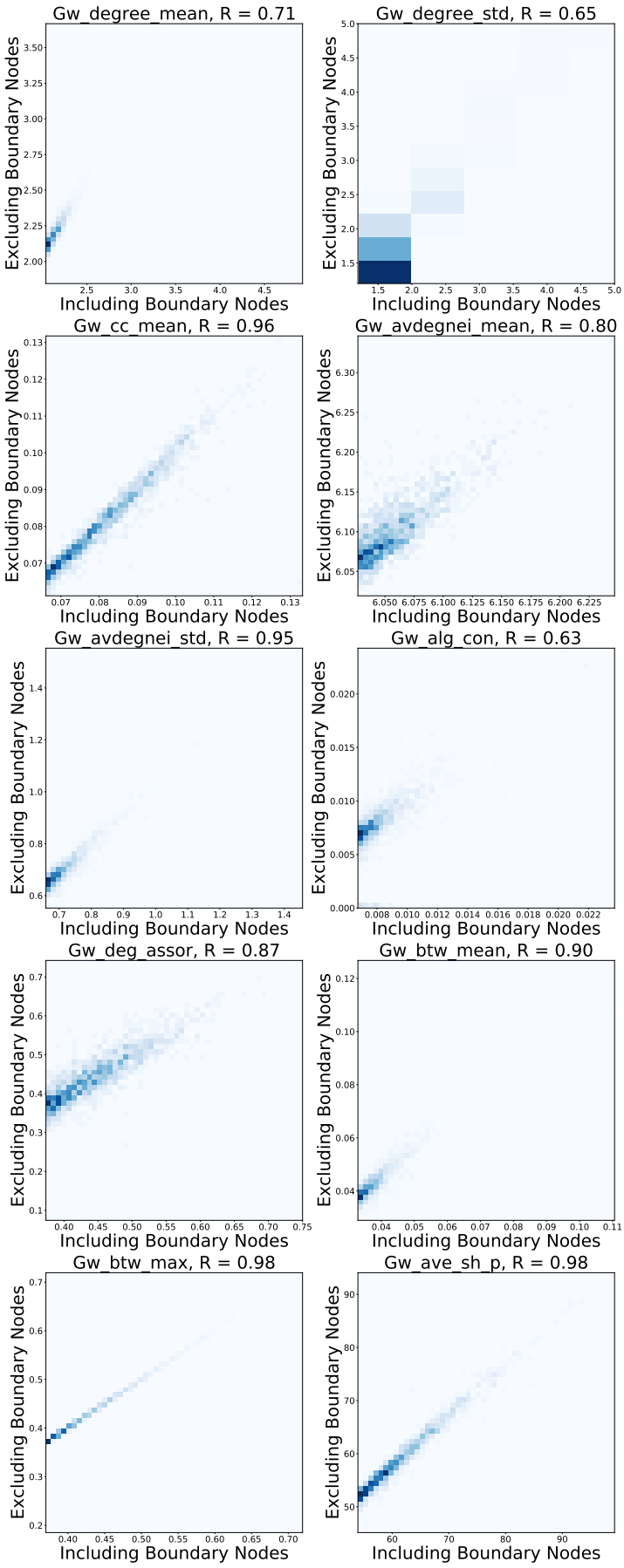


Figure A.30: The two-dimensional histograms for the 4^{th} quartile of the important network metrics. x-axis: network metric with the nodes located on boundaries, y-axis: network metric without the nodes located on boundaries

Figure A.31: Correlation heat-map of network metrics (without the nodes located on boundaries) with the first four principal components of Martin's metrics.

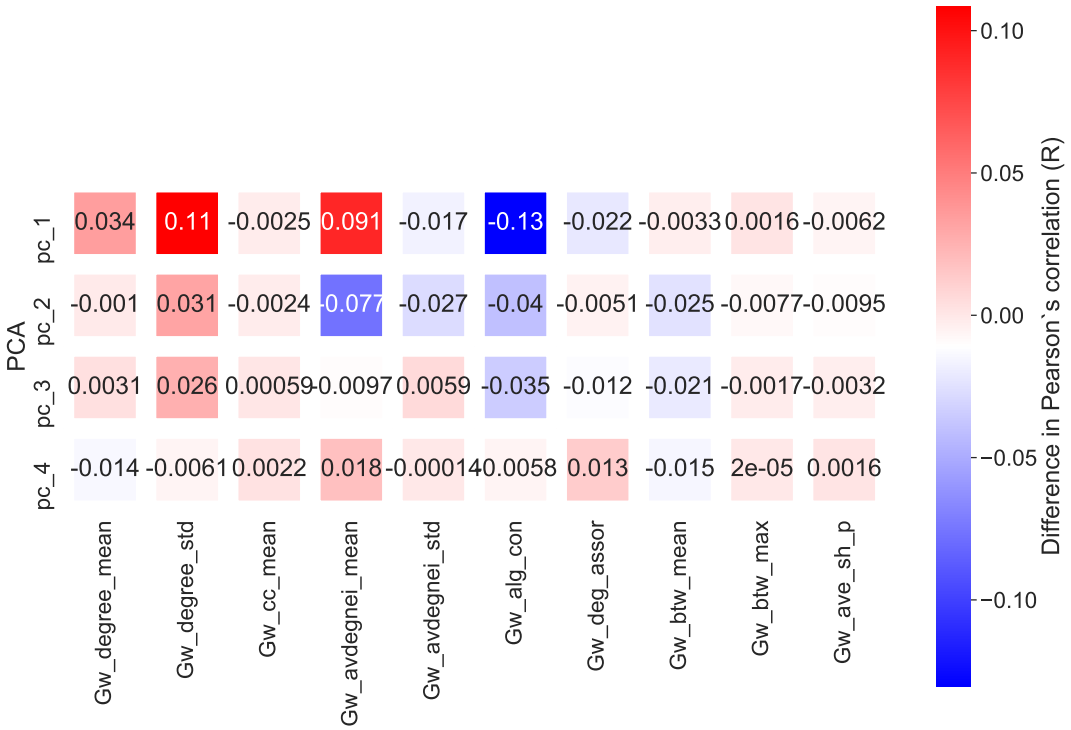


Figure A.32: The element-wise difference between two correlation heat-maps Figure A.31 and Figure 4.13

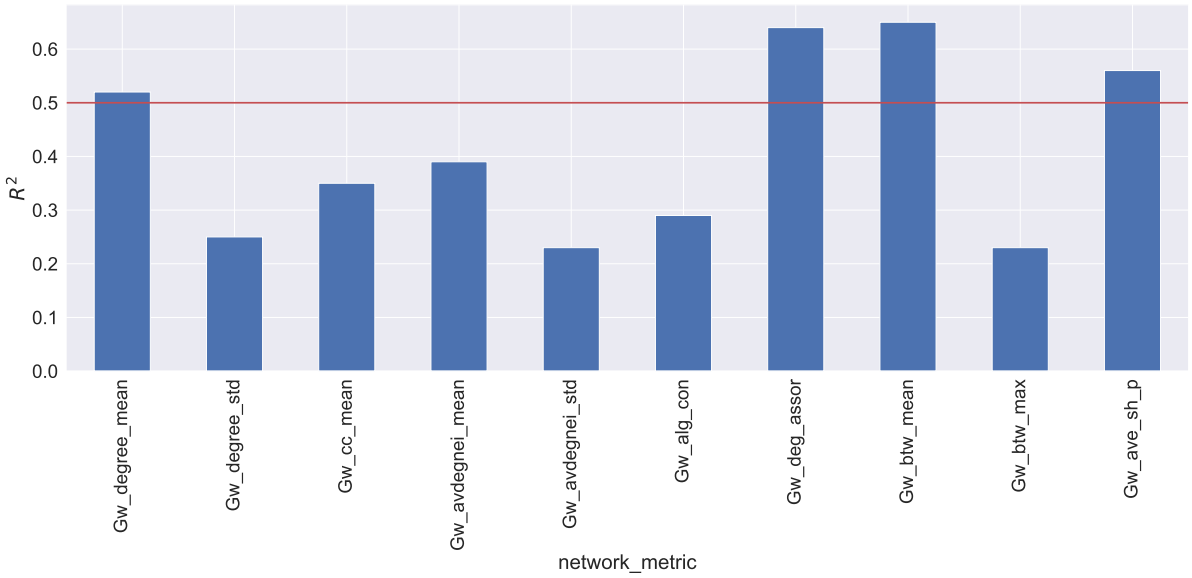


Figure A.33: The figure shows the R-squared values of each multiple linear regression model. In each model, the target feature is one of the network metrics (without the nodes located on boundaries) and the regressors are principal components of Martin's metrics.