GRAPH SPECTRAL DOMAIN BLIND WATERMARKING

Hiba Al-Khafaji and Charith Abhayaratne

Department of Electronic and Electrical Engineering, The University of Sheffield
Sheffield, S1 4ET, United Kingdom
Email: h.alkhafaji@sheffield.ac.uk, c.abhayaratne@sheffield.ac.uk

ABSTRACT

This paper proposes the first ever graph spectral domain blind watermarking algorithm. We explore the recently developed graph signal processing for spread-spectrum watermarking to authenticate the data recorded on non-Cartesian grids, such as sensor data, 3D point clouds, Lidar scans and mesh data. The choice of coefficients for embedding the watermark is driven by the model for minimisation embedding distortion and the robustness model. The distortion minimisation model is proposed to reduce the watermarking distortion by establishing the relationship between the error distortion using mean square error and the selected Graph Fourier coefficients to embed the watermark. The robustness model is proposed to improve the watermarking robustness against the attacks by establishing the relationship between the watermark extraction and the effect of the attacks, namely, additive noise and nodes data deletion. The proposed models were verified by the experimental results.

Index Terms— Graph spectral domain blind watermarking, Graph Fourier Transform (GFT), robustness, distortion.

1. INTRODUCTION

The most common approaches to protect graph-type data are: adding new nodes [1]; inserting extra edges [2] and embedding sub-graph [3]. Since these approaches are based on the node domain, they are not robust to many attacks and not secure. On the other hand, watermarking using spread spectrum has proven to be an effective approach in image protection, mainly due to advances in signal transforms. Our previous work [4] explored the first ever graph spectral domain non-blind watermarking and has proven to be very successful in the protection and authentication of the unstructured data. This paper proposes the first-ever graph spectral domain blind watermarking algorithm for authentication, which is a useful approach where the original signal is not available in the watermark extraction process.

For any watermarking system, the basic requirements are low error distortion and high robustness. The existing node-domain graph watermarking methods are also focused on the watermarking robustness against the attacks [5–10] as well as minimising the distortion [11–13]. Similarly in our previous work, we have proposed models minimising the embedding distortion [14] and making robust for scalable decoding attacks [15, 16] for general spread spectrum watermarking.

In this paper, within the proposed graph spectral domain blind watermarking, we propose a new model for choosing the embedding coefficients for minimising the embedding distortion and another model for choosing the embedding coefficients that are robust for attacks. In the distortion minimisation model, we need to establish the relationship between the error distortion metric and the selected coefficients to be watermarked in order to reduce the embedding distortion. The robustness model is proposed to improve the robustness in a way that the watermark can extract accurately after the attacks by establishing the relationship between the extraction process and the effect of the attack. We considered two types of attacks in this paper: additive noise and nodes data deletion. Finally, we combine the conditions of the two proposed models in order to satisfy the two main requirements of the watermarking. The main contributions of the proposed work are:

1. Proposal for a distortion minimisation model for graph spectral domain blind watermarking.
2. Proposal of a robustness model for graph spectral domain blind watermarking.

The rest of the paper is organized as follows: Section 2 presents the proposed method, followed by the performance evaluation in Section 3 and conclusions in Section 4.

2. THE PROPOSED METHOD

2.1. Graph Fourier Transform

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$, is an undirected graph without self-loops and multiple links between nodes, where $\mathcal{V}$ is the set of N vertices, $\mathcal{E}$ is the set of edges and $\mathbf{A}$ is the adjacency matrix with edge weights. We define the weight, $\mathbf{A}_{i,j}$ corresponding to an edge, $e_{i,j}$ connecting vertices i and j is as follows:

$$\mathbf{A}_{i,j} = \begin{cases} 1, & \text{if there is an edge } \ e_{i,j}, \\ 0, & \text{otherwise .} \end{cases} \tag{1}$$

We define the signal $\mathbf{x} : \mathcal{V} \rightarrow \mathbb{R}$. The combinatorial graph Laplacian matrix, $\mathbf{L}$, is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal matrix of vertex degrees, whose diagonal components are computed as follows:

$$\mathbf{D}_{(i,i)} = \sum_{j=0}^{N-1} \mathbf{A}_{(i,j)}, \qquad i = 0, 1, ..., N-1. \tag{2}$$

Since, $\mathbf{L}$, is a symmetric positive semidefinite matrix, from spectral projection theorem, there exists a real unitary matrix, $\mathbf{U}$, that diagonalizes $\mathbf{L}$, such that $\mathbf{U}\mathbf{L}\mathbf{U}^t = \Lambda = diag\{\lambda_\ell\}$ is a non-negative diagonal matrix , leading to an eigenvalue decomposition of $\mathbf{L}$ matrix as follows:

$$\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^t = \sum_{\ell=0}^{N-1} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^t, \tag{3}$$

where $\mathbf{u}_\ell$, the column vectors of $\mathbf{U}$, are the set of orthonormal eigenvectors of $\mathbf{L}$ with corresponding eigenvalues, $0 = \lambda_0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_{N-1} = \lambda_{max}$. [17]. The eigenvectors have been used in analysing graph spectra both algebraic and analytic wise [18]. The Graph Fourier Transform (GFT) and its inverse are defined as follows [17, 19]:

$$\mathbf{X}(\ell) = \sum_{i=0}^{N-1} \mathbf{x}(i)\mathbf{u}_\ell(i). \tag{4}$$

$$\mathbf{x}(i) = \sum_{\ell=0}^{N-1} \mathbf{X}(\ell)\mathbf{u}_\ell^t(i). \tag{5}$$

2.2. GFT domain blind watermarking

Firstly, the graph Fourier coefficients are calculated as in Eq. (4) and sorted in descending order to get the sorted coefficients, $\mathbf{X}_s(m)$. Then, a non-overlapping 3×1 running window is passed through the sorted GFT coefficients to embed the watermark in the median coefficient at each sliding position, as follows:

$$\mathbf{X}_{s_w}(m) = \lfloor \frac{\mathbf{X}_s(m-1) + \mathbf{X}_s(m+1)}{2} \rfloor + w, \tag{6}$$

where $\mathbf{X}_{s_w}$ is the watermarked coefficient, $\lfloor \mathbf{X} \rfloor$ denotes rounding of $\mathbf{X}$ to the largest integer smaller than $\mathbf{X}$ and $w > 0$ is the watermark information. In order for lossless extraction we restrict embedding for any 3 coefficients, if and only if it satisfies the condition: $\mathbf{X}_s(m-1) \geq \mathbf{X}_{s_w}(m) \geq \mathbf{X}_s(m+1)$.

For the watermark extraction, the GFT is performed on the watermarked graph signal, followed by sorting in descending order, to get sorted watermarked GFT coefficients, $\mathbf{X}_w(m)$. Then the watermark from each 3×1 running window with coefficients, $\mathbf{X}_w(m-1) \geq \mathbf{X}_w(m) \geq \mathbf{X}_w(m+1)$, is extracted as follows:

$$w' = \mathbf{X}_w(\ell) - \lfloor \frac{\mathbf{X}_w(\ell-1) + \mathbf{X}_w(\ell+1)}{2} \rfloor. \tag{7}$$

Let w_0 and w_1 are the chosen watermark values for embedding a 0 and 1 , respectively. The extracted watermark bit b' is determined based on a threshold T, where $T = (w_0 + w_1)/2$, as:

$$b' = \begin{cases} 0 \;\; , & \text{if } w' < T, \\ 1 \;\; , & \text{if } w' > T. \end{cases} \tag{8}$$

2.3. Embedding distortion minimisation

In order to establish the relationship between the error distortion using mean square error (μ) and the selected GFT coefficients for watermarking, we define mean square error (μ) in vertex domain between the original graph signal $\mathbf{x}$ and watermarked graph signal $\mathbf{x}_w$ as follows:

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} (\mathbf{x}(i) - \mathbf{x}_w(i))^2. \tag{9}$$

Since the GFT forms an orthogonal set of eigenvectors, according to the Parseval's Theorem, $\|\mathbf{x}\|^2 = \|\mathbf{X}\|^2$, where $\mathbf{x}$ is the graph signal in vertex domain and $\mathbf{X}$ is the GFT coefficient [19]. Since the GFT is orthonormal, we can extend this to the sum of the error power in the input graph signal, $\Delta\mathbf{x}$, and to the sum of the error power in the GFT domain $\Delta\mathbf{X}$ as follows:

$$\sum_i |\Delta\mathbf{x}(i)|^2 = \sum_\ell |\Delta\mathbf{X}(\ell)|^2. \tag{10}$$

From Eq. (9) and Eq. (10), we get

$$\mu = \frac{1}{N} \sum_\ell |\Delta\mathbf{X}(\ell)|^2. \tag{11}$$

From Eq. (6), we can estimate each $\Delta\mathbf{X}(\ell)$ as

$$\Delta\mathbf{X}_s(m) = \lfloor \frac{\mathbf{X}_s(m-1) + \mathbf{X}_s(m+1)}{2} \rfloor + w - \mathbf{X}_s(m), \tag{12}$$

Thereby leading to

$$\mu \propto \sum (\lfloor \frac{\mathbf{X}_s(m-1) + \mathbf{X}_s(m+1)}{2} \rfloor - \mathbf{X}_s(m))^2. \tag{13}$$

Therefore to minimise μ, for each embedding coefficient triple, $\lfloor 0.5(\mathbf{X}_s(m-1) + \mathbf{X}_s(m+1)) \rfloor - \mathbf{X}_s(m)$ must be close to 0 or in other words, the gradient difference, $[(\mathbf{X}_s(m-1) - \mathbf{X}_s(m)) - (\mathbf{X}_s(m) - \mathbf{X}_s(m+1))]$ must be close to 0.

2.4. On enhancing robustness

The aim of the proposed model is to find the GFT coefficients values which are capable of extracting the watermark after the attack in GFT domain. At this point, we consider two types of attacks: additive noise and deletion of random nodes data on test graphs. The watermarked GFT coefficients values $\mathbf{X}_w(\ell)$

are changed based on the value of the modification due to attack Δ_a as follows:

$$\mathbf{X}'_w(\ell) = \mathbf{X}_w(\ell) + \Delta_a, \tag{14}$$

where $\mathbf{X}'_w(\ell)$ are the watermarked GFT coefficients values after the attack. The modification value Δ_a depends on the type of attack. For example, the modification value of deleting nodes data depends on the number of the node data which are deleting and their positions in the graph.

To extract the watermark information w' after the attack, we have new GFT coefficients values $\mathbf{X}'_w(\ell'-1), \mathbf{X}'_w(\ell')$ and $\mathbf{X}'_w(\ell'+1)$:

$$w' = \mathbf{X}'_w(\ell') - \lfloor \frac{\mathbf{X}'_w(\ell'-1) + \mathbf{X}'_w(\ell'+1)}{2} \rfloor. \tag{15}$$

At this point, three cases of the watermark bits are considered: embedding only '0' bits, embedding only '1' bits and embedding '0' and '1' bits.

Embed '1': To extract the correct watermark after embedding '1' bits, the watermarked coefficients should be in the range:

$$\lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + T \leq X'_w(\ell') < \lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + w_1. \tag{16}$$

Embed '0': For embedding '0' bits, we can detect the correct watermark bits when the watermarked coefficients are in the range:

$$\lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + w_0 \leq X'_w(\ell') < \lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + T. \tag{17}$$

Embed '0' and '1': By combining the two cases above, we can find the condition of correct detection of the watermark bits when embedding '0' and '1'. The range of the GFT coefficients which retain the watermark bits correctly is:

$$\lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + w_0 \leq X'_w(\ell') < \lfloor \frac{X'_w(\ell'-1) + X'_w(\ell'+1)}{2} \rfloor + w_1. \tag{18}$$

Fig. 1 shows the range of the GFT coefficients capable of retaining the watermark bits after the attacks.

2.5. Joint robust-low distortion blind watermarking

To satisfy the two complementary requirements of the graph watermarking, we combined the two proposed models for selecting the GFT coefficients for watermark embedding. Eq. (18) and Eq. (13) are used for meeting the robustness for given maximum robustness followed by choosing coefficient triple that has the gradient difference close to 0 for minimising the distortion.

3. PERFORMANCE EVALUATION

The proposed GFT domain blind watermarking algorithm with the embedding distortion minimisation and robustness models was tested using the graph watermarking dataset [20].

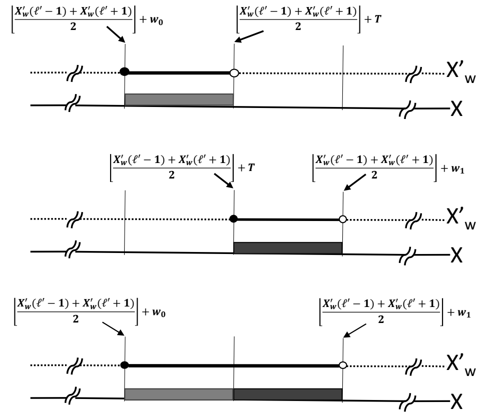
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + w_0$
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + T$
X'w
X
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + T$
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + w_1$
X'w
X
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + w_0$
$\left\lfloor\frac{X'_w(\ell'-1)+X'_w(\ell'+1)}{2}\right\rfloor + w_1$
X'w
X

Fig. 1: The range of the GFT coefficients capable of extracting the watermark bits correctly, Row 1: Embedding only $w =' 0'$, Row 2: Embedding only $w =' 1'$, Row 3: Embedding $w =' 0'$ and $'1'$.

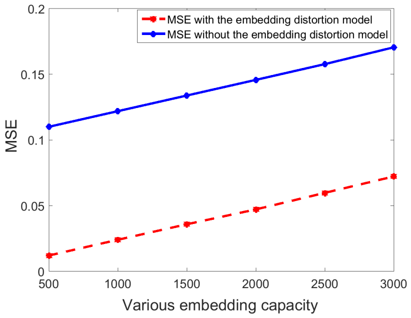

Fig. 2: Embedding distortion performance

3.1. Embedding distortion performance

The embedding distortion was measured using the embedding model and without using the model for the proposed blind GFT watermarking method, when the same number of watermarking bits were embedded. As shown in Fig. 2, the proposed model provides lower distortion using the proposed model. It can be observed that the embedding distortion is increased when the embedding capacity is increased for both the methods.

3.2. Robustness performance

The robustness model was verified by the experimental results by comparing the Hamming distance (HD) of the extracted watermark after two types of attacks: additive noise and deleting random nodes data using the original blind algorithm (without using the model) and the algorithm with using the robustness model by choosing the GFT coefficients which satisfy the conditions (in Eq. (16),Eq. (17), and Eq. (18)) for 3 embedding scenarios. We considered the pseudo-random

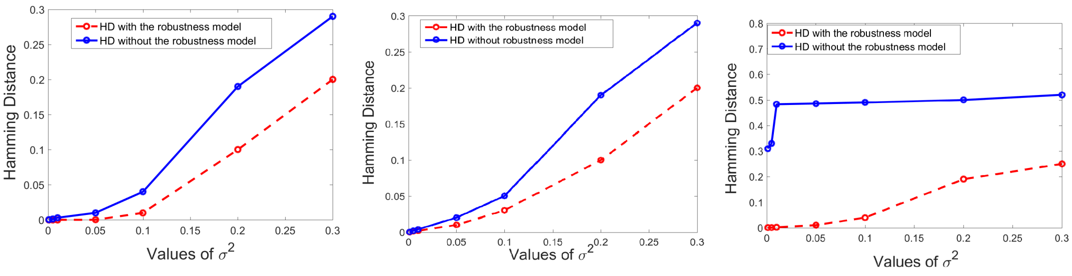

Fig. 3: The average values of Hamming distance (HD) after additive noise for various σ^2 values: Column 1: Embedding ’b=1’. Column 2: Embedding ’b=0’.Column 3: Embedding ’0’ and ’1’ .

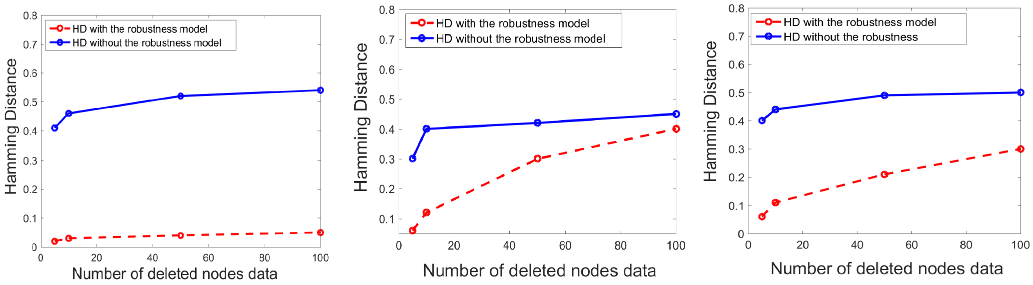

Fig. 4: The average values of Hamming distance (HD) after deleting different number of random nodes data: Column 1: Embedding ’b=1’. Column 2: Embedding ’b=0’. Column 3: Embedding ’0’ and ’1’ .

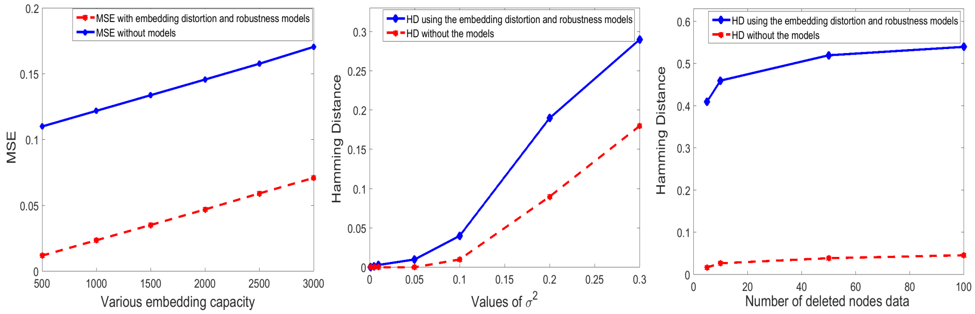

Fig. 5: Combining the embedding distortion minimisation and robustness models. Column 1: Embedding distortion. Column 2: Robustness to additive noise, embedding ’b=1’. Column 3: Robustness to delete random nodes data, embedding ’b=1’.

binary sequences as a watermark with three cases: b = 1, b = 0 and b = 0 and 1 are embedded in the GFT coefficients of the sensor graph. It can be observed that the Hamming Distance (HD) was reduced when using the proposed model, this mainly means the robustness is improved by using the proposed model for various σ^2 values and deleting a different number of nodes data respectively as shown in Fig. 3 and Fig. 4. Also, we can notice that the effect of the attack decreases by using the model for example in additive noise we can see that the Hamming Distance is zero when $\sigma^2 < 0.05$, this mainly means there is no effect of the noise in this case and the watermark information can extract accurately. By combining the two models we obtained a watermarking approach with low distortion and robust to attacks as shown in Fig. 5.

4. CONCLUSIONS

In this paper, we have proposed GFT domain blind watermarking. The proposed approach includes two novel models for minimising the embedding distortion on host graph data and for making the watermarks robust for attacks, namely, noise and node deletion. The distortion minimisation model requires to choose the sorted coefficient triples with the gradient difference close to 0 to minimise the distortion, while the robustness model is designed to improve the robustness against the attacks based on selecting the GFT coefficients which satisfy the specific conditions for watermark embedding. The proposed models are supported by experimental evaluation, which shows the benefit of using both models for GFT domain blind watermarking.

5. REFERENCES

[1] A. Saha, D. Bhaumik, and S. Pathak, "Signature hiding and recovery in a graph coloring solutions using modified genetic algorithm," in *Proc. of International Conference on Computer and Information Technology (ICCIT)*, 2011, pp. 50–55.

[2] G. Qu and M. Potkonjak, "Analysis of watermarking techniques for graph coloring problem," in *Proc. of the International conference on Computer-aided design*, 1998, pp. 190–193.

[3] X. Zhao, Q. Liu, H. Zheng, and B. Y Zhao, "Towards graph watermarks," in *Proc. of the Conference on Online Social Networks*, 2015, pp. 101–112.

[4] H. Al-khafaji and C. Abhayaratne, "Graph spectral domain watermarking for unstructured data from sensor networks," in *Proc. of International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[5] J-W. Cho, R. Prost, and H-Y. Jung, "An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 142–155, 2007.

[6] R. Darazi, R. Hu, and B. Macq, "Applying spread transform dither modulation for 3D-mesh watermarking by using perceptual models," in *Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 1742–1745.

[7] J-S. Tsai, J-T. Hsiao, W-B. Huang, and Y-H. Kuo, "Geodesic-based robust blind watermarking method for three-dimensional mesh animation by using mesh segmentation and vertex trajectory," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 1757–1760.

[8] X. Rolland-Neviere and P. Doerr, G.and Alliez, "Security analysis of radial-based 3d watermarking systems," in *International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2014, pp. 30–35.

[9] X. Rolland-Neviere, G. Doërr, and P. Alliez, "Anticropping blind resynchronization for 3D watermarking," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1702–1706.

[10] J-U. Hou, D-G. Kim, and H-K. Lee, "Blind 3D mesh watermarking for 3D printed model by analyzing layering artifact," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2712–2725, 2017.

[11] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3-D mesh quality assessment," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 247–256, 2007.

[12] N. Medimegh, S. Belaid, M. Atri, and N. Werghi, "3D mesh watermarking using salient points," *Multimedia Tools and Applications*, pp. 1–23, 2018.

[13] S. Borah and B. Borah, "Quantization index modulation (QIM) based watermarking techniques for 3D meshes," in *Proc. of International Conference on Image Information Processing (ICIIP)*. IEEE, 2017, pp. 1–6.

[14] D. Bhowmik and C. Abhayaratne, "A generalised model for distortion performance analysis of wavelet based watermarking," in *International Workshop on Digital Watermarking*, 2008, pp. 363–378.

[15] D. Bhowmik and C. Abhayaratne, "Quality scalability aware watermarking for visual content," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5158–5172, 2016.

[16] C. Abhayaratne and D. Bhowmik, "Scalable watermark extraction for real-time authentication of JPEG2000 images," *Journal of real-time image processing*, vol. 8, no. 3, pp. 307–325, 2013.

[17] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[18] S. Butler, "Algebraic aspects of the normalized laplacian," in *Recent Trends in Combinatorics*, pp. 295–315. 2016.

[19] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[20] H. Al-khafaji and C. Abhayaratne, "Graph watermarking dataset," `https://figshare.shef.ac.uk/s/3d92615f2a87f0ea719e`, 2019, Accessed: 18-02-2019.