ROOM GEOMETRY ESTIMATION FROM ACOUSTIC ECHOES USING GRAPH-BASED ECHO LABELING

Ingmar Jager, Richard Heusdens and Nikolay D. Gaubitch

Delft University of Technology

ABSTRACT

A computer being able to estimate the geometry of a room could benefit applications such as auralization, robot navigation, virtual reality and teleconferencing. When estimating the geometry of a room using multiple microphones, the main challenge is to identify which reflections, or echoes, originate from the same wall and can, therefore, be modeled by a virtual source outside the room using the mirror image source model. In this paper we present a new and efficient method to disambiguate the echoes using a graph theoretical approach where echo combinations are modeled as nodes in a graph and the problem is stated as a maximum independent set problem. Once the echoes are correctly labelled, we know the locations of the virtual sources from which we can infer the room geometry. Experiments for shoe-box shaped rooms show that we can reliably estimate the room geometry within seconds on contemporary hardware and achieve centimeter precision on finding the vertices of the room.

Index Terms— room geometry estimation, mirror image source model, independent sets

1. INTRODUCTION

Recently there has been an increasing interest in developing computer algorithms that can measure the shape of a room using acoustic echoes. This can benefit several applications. In auralization one needs to model the source, receiver and transmission medium, of which the latter can be accurately modeled only if information about the room geometry is available. For teleconferencing one may want to take into account the reflections of sound, also called reverberation, prior to the excitation of a sound signal. Robot navigation benefits from it in in the sense that it aids to avoid unsafe conditions and dangerous situations such as collisions.

Reflections, or echoes, contain information about the geometry of the room. The reflections can be modeled by virtual sources using the mirror image source model [1], and finding the virtual sources implies finding the surfaces in a room. When using multiple microphones, the echoes from each wall do not necessarily arrive at each microphone in the same order. The main challenge is to disambiguate the echoes and label them according to the wall they originate from.

Several methods for obtaining the room shape from acoustic echoes have been proposed. In [2], a 2D room shape is found by identifying echoes in the room impulse response from a source to a single microphone, whereas in [3] the 2D room shape is found using multiple sources and microphones. In [4] a method for estimating 3D room shapes is presented, where the arrangement of microphones is kept small enough so that we can assume that echoes will cluster together in time. Its applicability, however, is restricted due to the many restrictions on the microphone and source locations. In [5], Dokmanić et al. describe a method for room shape estimation

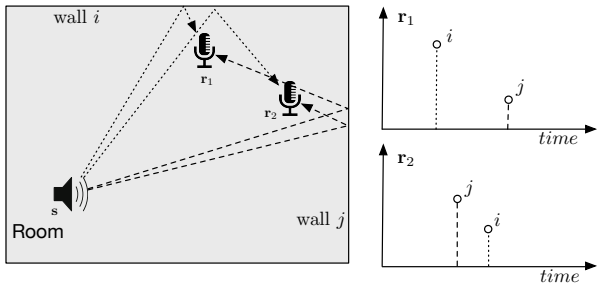


Fig. 1: Illustration of different order of arrival of wall reflections.

by recording the echoes of a single source using five microphones which can be placed arbitrarily in the room. They utilize the properties of Euclidean distance matrices and use multidimensional scaling to iteratively find the room shape. Although this method can be very accurate, it is computationally expensive and, in particular for realistic 3D scenarios, only suitable for off-line calculations.

In this paper we propose a high-accuracy efficient method to obtain the room shape in 3D scenarios. The proposed algorithm uses five microphones and $N \geq 2$ sources (or one single source excited at different locations), and uses a graph theoretical approach where echo combinations are modeled as nodes and the problem is stated as finding the maximum independent set in the graph. In this paper we will restrict ourselves to shoe-box shaped rooms, although the method itself can be used for arbitrary room shapes.

2. LOCALIZATION OF IMAGE SOURCES

In order to estimate the geometry (walls) of a shoe-box shaped room, we need to locate both the source and the six virtual sources corresponding to the first-order reflections of the walls, floor, and ceiling. These locations can be computed using trilateration once we know the distance between the different sources and the microphones, where the distances itself can be obtained from the time-of-arrivals (TOAs) of the different reflections. Although the estimation of the TOAs, and therefore the distances, is straightforward, it is not clear which echo originates from which wall since reflections may arrive in different order at the microphones. This ambiguity in the echoes is illustrated in Figure 1. In order to correctly identify which echo originates from which wall, we follow the approach of [5], which makes use of the properties of a Euclidean distance matrix (EDM). The EDM of a set of N points (locations), say $x_1, \ldots, x_N$, is an $N \times N$ matrix containing all squared Euclidean distances between the points of consideration. That is, entry (i, j) of the EDM is given by $\|x_i - x_j\|^2$, and as a consequence, its diagonal entries are

zero. It can be shown [6] that the rank of an EDM, for points in three-dimensional space, must have a rank less than or equal to five.

Consider the 3D point set containing the known (relative) positions of the M microphones, say $r_1, \ldots, r_M$. Using this set, we can construct an EDM $R \in \mathbb{R}^{M\times M}$, which has, assuming $M \geq 5$ and microphones are not co-located, rank five. If we would augment this matrix by adding the distances from the M microphones to one particular virtual source, resulting in an augmented EDM $\tilde{R}$, the rank of the augmented matrix will not increase. If we augment R by distances to different virtual sources, however, the augmented matrix is *not* an EDM anymore and the rank of $\tilde{R}$ will be larger than five. Hence, a brute force method for identifying which echo originates from which wall is to try out every possible echo combination and check whether a certain combination gives rise to an EDM. Except from the fact that finding the correct echo combination is a NP-hard problem, the method breaks down in case there is measurement noise in the TOAs, which will always be the case in any practical application.

In [5] this problem is tackled using multidimensional scaling (MDS) by iteratively finding the best matching EDM. However, the algorithm, although being asymptotically optimal, is computationally expensive and not suitable for real-time applications like robot navigation. In the following section we will present an alternative method to reduce the computational complexity of finding the correct echo combination. We first exclude some echo combinations based on the singular values of the augmented EDM, after which the remaining echo combinations are modeled as nodes in a graph and the problem is formulated as a maximum independent set problem.

2.1. Complexity reduction

In order to reduce the search space of possible echo combinations, we will first exclude echo combinations based on the singular values of the augmented EDM $\tilde{R}$. In the case we augmented R by a correct echo combination, we have that $\mathrm{rank}(\tilde{R}) = 5$, which implies that the SVD of $\tilde{R}$ contains only five nonzero singular values. If, on the other hand, we used an incorrect echo combination or there were measurement errors in the TOAs, we have $\mathrm{rank}(\tilde{R}) > 5$. As a consequence, excluding all augmented EDMs having a rank larger than five would exclude correct echo combinations in the presence of measurement noise. To overcome this possible problem, we note that if we perturb a rank-5 matrix by adding a distortion matrix, say D, then $\sigma_6 \leq \|D\|_2$ [7, Theorem 2.5.3], where σ_6 denotes the sixth largest singular value of $\tilde{R}$. Hence, noise has a definable effect on our ability to detect rank; if the singular values are larger than $\|D\|_2$ we know they did not just come from the noise. Therefore, instead of considering the rank of the augmented matrix, we will consider the ϵ-rank [7] of $\tilde{R}$, which is defined as

$$\mathrm{rank}(\tilde{R}, \epsilon) = \min_{\|\tilde{R}-X\|_2 \leq \epsilon} \mathrm{rank}(X).$$

As a consequence, by excluding all $\tilde{R}$ having a ϵ-rank larger than five, we will exclude echo combinations that give rise to an augmented EDM having more than five singular values larger than a tolerance ϵ. Since the tolerance can not be set arbitrarily low, there will be false positives which need to be excluded in a second stage.

Given our reduced set of candidate echo combinations, say $C_\epsilon = \{c_i\},\ c_i \in \mathbb{R}^M$, we need to find a subset of six elements c_i, each containing the distances from one particular image source to all the M receivers. A key observation we can make here is that these six vectors are very unlikely to have elements in common. This will only happen by special construction and means that different first-order image sources have identical distance to a particular receiver. As a consequence, we can reduce our search space to the set of c_is that have no elements in common.

To illustrate this, we consider the following 2D toy example in which there are three microphones and four walls. Assuming we can identify the first-order reflections, we find four distances for each of the three receivers. Next, assume that after applying the ϵ-rank test, the (reduced) set of possible echo combinations is given by

$$C_\epsilon = \begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \end{array} \begin{array}{c} \begin{array}{cccccccc} c_0 & c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 \end{array} \\ \left[\begin{array}{cccccccc} 10 & 11 & 11 & 11 & 12 & 12 & 14 & 14 \\ 20 & 20 & 21 & 22 & 22 & 21 & 24 & 24 \\ 30 & 31 & 31 & 31 & 33 & 31 & 31 & 34 \end{array}\right] \end{array}. \quad (1)$$

Since there are four walls, we need to find four candidate echo combinations $c_i \in C_\epsilon$ that have unique distances to the three receivers. By inspection of (1), we conclude that the subset $(c_0,\, c_2,\, c_4,\, c_7)$ is the required subset since this is the only four-element combination of c_is that do not have common elements; the other elements (c_1, c_3, c_5, c_6) are spurious echo combinations that happened to pass the rank test. Note that in general, there can be more than one four-element set (or, in the 3D shoe box room case, six-element set) that have no elements in common. In that case we need to find all these subsets and we have to decide afterwards which one is the correct one. In the next subsection we will describe an efficient method for finding the required set(s) based on graph theory.

2.1.1. Independent sets

A *graph* is an abstract representation of a set of objects where some pairs of the objects are connected by links. The interconnected objects are represented by mathematical abstractions called *vertices* or *nodes*, and the links that connect some pairs of vertices are called *edges*. A graph, denoted by G, thus consists of some finite number, say n, of nodes, which will be labeled and represented as a vertex set $V = \{1, \ldots, n\}$, and edges representing connections between the nodes represented by $E \subseteq V \times V$. We write $G = (V, E)$. Typically, a graph is depicted in diagrammatic form as a set of dots for the vertices, joined by lines or curves for the edges (see Figure 2 for an example). In this paper we will focus on *undirected* graphs only.

Let each candidate echo combination $c_i \in C_\epsilon$ be represented by a node $i \in V$, where $|V|$, the cardinality of the vertex set, is equal to the number of echo combinations that passes the ϵ-rank test. For every two candidates $c_i, c_j \in C_\epsilon$ that have one or more elements in common, we define an edge in E. Figure 2 illustrates the above definitions using the toy example presented in the previous section. Hence, there a eight nodes in total where each node i represents echo combination c_i. In this example, echo combination c_0 has one element in common with echo combination c_1, and c_2 has two elements in common with c_1, and so on, whereas there are no common elements between the echo combinations c_0 and c_2. By inspection of Figure 2 we conclude that the node set $\{0, 2, 4, 7\}$ (indicated by the blue-colored nodes) is a set of vertices no two of which are adjacent. That is, it is the set of vertices that do not have a direct interconnection, and is generally referred to as an *independent set* [8]. There can, however, be many independent sets in a graph. By definition, each subset of an independent set is also an independent set. A *maximal* independent set is an independent set such that it is not possible to include any other node from the graph in the set without it ceasing to be an independent set. For the problem at hand, we have to find the independent set of largest possible size, which is referred to as the *maximum* independent set. In our toy example, this size is four,

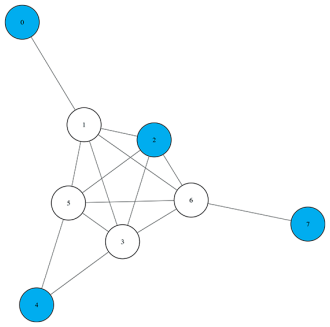


Fig. 2: Representing echo combinations as nodes in a graph. The set of blue nodes correspond to the (maximum) independent set.

and we conclude that the independent set $\{0, 2, 4, 7\}$ is a maximum independent set and, therefore, solves our problem.

Finding a maximum independent set is an NP-hard problem. It can, however, be solved more efficiently than the trivial $\mathcal{O}(2^n)$ bound given by a naive brute force method. For example, [9] presented an exponential space algorithm using time $\mathcal{O}(2^{0.276n})$ to solve the problem. Since there can be more than one maximum independent set in a graph, we need to find all maximum independent sets. One way to do this is solving the so-called maximal independent set listing problem [10]. Listing all maximal independent sets also yields all maximum independent sets, since a maximum independent set is a maximal independent set by definition.

2.2. Finding the correct echo locations

Let $S_{\max}$ denote the set of maximum independent sets in a graph. In terms of the room geometry estimation problem, each maximum independent set contains echo combination that do not have elements in common. If $S_{\max}$ contains only one element, we have found the correct echo combination from which we can infer the room geometry. If, however, $S_{\max}$ contains more than one such set, we have to decide which one is the correct one. In this paper we propose to use the source localization algorithm proposed by Pollefeys [11] which gives, given the distances between sources and receivers, the location of both sources and receivers up to a unitary transform (rotation, reflection) and translation.

Pollefeys' method requires at least ten sources and five microphones. Using the image source method, we can model reflections as virtual sources and we conclude that we need to do at least two measurements, yielding two real sources and 12 image sources. Given $N = 2$ sources, we construct the input data for the Pollefeys method as

$$\Delta = \begin{bmatrix} S & E_1 & E_2 \end{bmatrix}, \tag{2}$$

where $S \in \mathbb{R}^{M\times N}$ contains the distances from the M microphones to the N real sources, and $E_1, E_2 \in \mathbb{R}^{M\times 6}$ contain distances from the microphones to six echo combinations originating from source 1 and 2, respectively, found by the maximum independent set algorithm.

The Pollefeys method gives us the coordinates of the receivers and sources given that the input data corresponds to the correct combination of echoes. If the echo combination is not correct, the estimated coordinates will, in general, be completely wrong. Since we know the pairwise distances between the microphones, we can compare the estimated microphone locations using Procrustes alignment [12] with the true microphone locations to find out whether the input data was correct or not. If the receiver locations turn out to be correct, the (image) source coordinates will be correct as well. As a consequence, we will try out all possible combinations E_1 and E_2 and select the one that gives the smallest receiver reconstruction error.

Average localization error	0.0235 m
Localization error variance	2.21×10^{-3}
Minimum localization error	1.08×10^{-3} m
Average run time	2.43 s
Run time variance	0.51
Minimum run time	1.35 s

Table 1: Average results for estimating room geometry for shoe-box shaped rooms

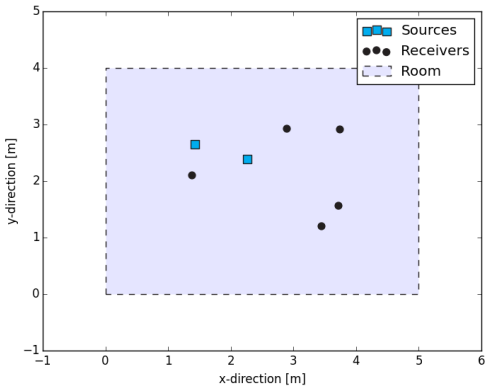
3. ROOM RECONSTRUCTION

In order to determine the room geometry, we first have to determine which echo originates to which wall. To do so, let s_i denote the location of source i and s_{ij} the location of the image source of source i with respect to wall j. With this the normal vector of wall j is given by $n_j = s_i - s_{ij}$ for all $i = 1, \ldots, N$. Given the estimated source locations obtained by the Pollefeys method, denoted by $\hat{s}_i$ and $\hat{s}_{ij}$, we can compute $\hat{s}_i - \hat{s}_{ij}$ for all i, j, and cluster the result into six 2-element sets using the k-means clustering algorithm [13, 14]. As a result, we have collected all image sources belonging to the same wall in one cluster, and we can estimate points on wall j as $w_{ij} = (\hat{s}_{ij} + \hat{s}_i)/2$ for every source i. Figure 3a shows an example shoe-box shaped room where we randomly placed $N = 2$ sources and $M = 5$ microphones. The blue squares indicate the source locations whereas the black dots indicate the microphone locations. Figure 3b shows the result of the procedure described above. The purple triangles show the estimated image source locations and the green circles the estimated wall points w_{ij}. Having the wall points estimated, the vertices of the room are simply found as the intersection of the lines through the different wall points.

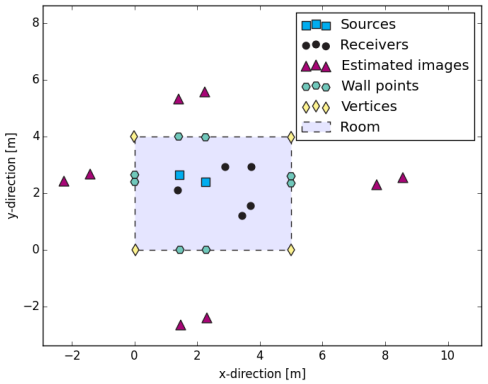
We can improve the accuracy of our room geometry estimation by including more than two sources. In general, when we use $N \geq 2$ sources, the input to Pollefeys' method is given by (2) where $E_1, E_2 \in \mathbb{R}^{M\times 6}$ contain distances from the M microphones to six echo combinations originating from source i and j, where $1 \leq i, j \leq N, i \neq j$. Since we can make $\binom{N}{2}$ such combinations, we get $2\binom{N}{2}$ estimated image sources in total, resulting in

$$\frac{2}{N}\binom{N}{2} = \frac{(N-1)!}{(N-2)!} = N - 1,$$

estimations per image source. Figure 4a shows the result in case we have $N = 5$ sources, resulting in four estimated locations per image source, and as a consequence, in $4N$ wall points per wall. The wall itself is obtained by a least-squares fit through the wall points. Figure 4b zooms in on the bottom part of the room depicted in Figure 4a. Obviously, adding more sources will improve the estimation performance but will increase the computational complexity as well. Depending on the application, a compromise needs to be found between performance gain and computational cost.

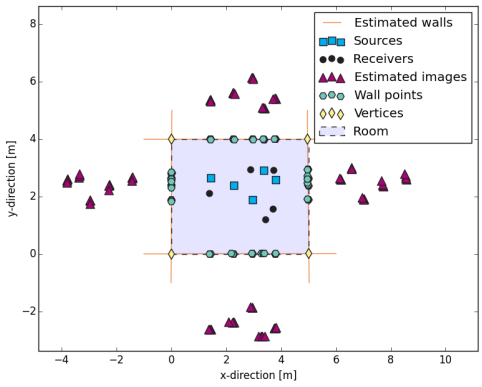


(a) Initial room setup

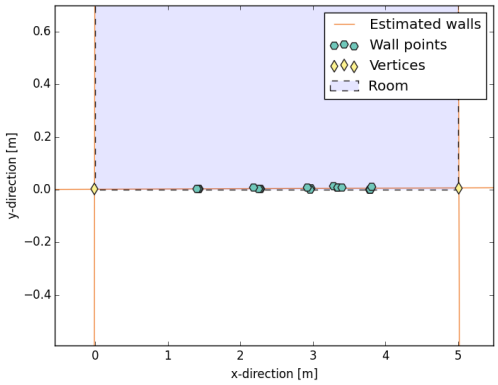


(b) Estimated room geometry

Fig. 3



(a) Estimated room geometry using $N = 5$ sources



(b) Zoomed-in version of the results shown in Figure 4a

Fig. 4

4. EXPERIMENTAL RESULTS

In this section we will present experimental results obtained by computer simulations for 3D shoe-box shaped rooms. To verify the proposed method, we generated rooms of 12 different dimensions having volumes in the range of $120 - 500$ m^3. For every room geometry, experiments were repeated 50 times. The experiments were done for $N = 5$[1] sources and $M = 5$ microphones, where the sample frequency was 96 kHz. For a particular setting, room impulse responses from sources to microphones were generated using room acoustics simulation software [15]. After estimating the room geometry, we measured the performance of the algorithm by computing the 2-norm of the location errors of the room vertices. Experiments were run on a MacBook Pro Mid 2012, 2.3 GHz Core i7 processor in Python 3.4.3 using Scipy [16], Numpy [17] and NetworkX [18]. Table 1 shows results averaged over all (600) experiments.

In order to compare these results to results obtained by the method described in [5], we observed that finding a particular set of six image sources (with comparable accuracy as presented in Table 1) for a particular source takes approximately one hour. We could, however, apply the same complexity reduction method as proposed in Section 2.1. By doing so, the computation time per set of image source was reduced to approximately 50 seconds. Note that the newly proposed algorithm takes about 2.4 seconds to find *all* sets of image sources which is a few orders of magnitude faster than state-of-the-art solutions.

[1]Informal tests showed that using five sources gave rise to an acceptable computational complexity given the location accuracy.

5. CONCLUSIONS

In this paper we considered the problem of estimating the room geometry using acoustic echoes. The proposed solution is based on jointly estimating the source and receiver locations of a set of candidate echo combinations, and compare the receiver location estimates thus obtained with the known (relative) receiver locations. In order to reduce the computational complexity of the proposed method, we first exclude some echo combinations based on the singular values of the augmented EDM using a ϵ-rank test, after which the remaining echo combinations are modeled as nodes in a graph and the problem is formulated as a maximum independent set problem. The proposed method uses $N \geq 2$ sources and five microphones to estimate the room geometry of a shoe-box shaped room. Experimental results obtained by computer simulation showed that the proposed algorithm estimates the vertices of the rooms with an average error of 2.35 cm within a few seconds (2.43 seconds on average), which is a few orders of magnitude faster than existing methods.

6. REFERENCES

[1] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[2] Ivan Dokmanić, Yue M Lu, and Martin Vetterli, "Can one hear the shape of a room: The 2-d polygonal case," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 321–324.

[3] Fabio Antonacci, Jason Filos, Mark RP Thomas, Emanuël AP Habets, Augusto Sarti, Patrick Naylor, Stefano Tubaro, et al., "Inference of room geometry from acoustic impulse responses," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 10, pp. 2683–2695, 2012.

[4] Sakari Tervo and Timo Tossavainen, "3d room geometry estimation from measured impulse responses.," in *ICASSP*, 2012, pp. 513–516.

[5] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12186–12191, 2013.

[6] John Clifford Gower, "Properties of euclidean and non-euclidean distance matrices," *Linear Algebra and its Applications*, vol. 67, pp. 81–97, 1985.

[7] G.H. Golub and C.F. Van Loan, *Matrix Computations*, North Oxford Academic, Oxford, third edition, 1983.

[8] R. Diessel, *Graph Theory*, vol. 173 of *Graduate Texts in Mathematics*, Springer-Verlag, Heidelberg, fourth edition, 2010.

[9] John Michael Robson, "Algorithms for maximum independent sets," *Journal of Algorithms*, vol. 7, no. 3, pp. 425–440, 1986.

[10] David Eppstein, "All maximal independent sets and dynamic dominance for sparse graphs," *ACM Trans. Algorithms*, vol. 5, no. 4, pp. 38:1–38:14, Nov. 2009.

[11] Marc Pollefeys and David Nister, "Direct computation of sound and microphone locations from time-difference-of-arrival data.," in *ICASSP*, 2008, pp. 2445–2448.

[12] J.C. Gower and G.B. Dijksterhuis, *Procrustes Problems*, Oxford University Press, Oxford, 2004.

[13] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–780, 1965.

[14] Leonard Kaufman and Peter J. Rousueeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[15] Andrew Wabnitz, Nicolas Epain, Craig Jin, and André van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, 2010.

[16] Eric Jones, Travis Oliphant, Pearu Peterson, et al., "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2015-08-22].

[17] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[18] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, Aug. 2008, pp. 11–15.