Food Chain Analysis Based on Graph Centrality Indicators

Victor Kolesnikov, Vasily Anikin, Ekaterina Mosolova, Alexey Faizliev, Sergei Mironov, Maria Zemlyanskaya, Michael Pleshakov and Sergei Sidorov
Saratov State University, 3, Astrakhanskaya str., Saratov, 410012 Russia
E-mail: kolvictor64@mail.ru

Abstract. The knowledge of the structure and complexity of the food web is central for a better understanding of the ecosystem functioning. The food web analysis examines both species and energy flows among them, providing a natural basis for characterizing the ecological roles of the species and the mechanisms by which biodiversity influences the dynamics of the ecosystem. This paper, for the first time, presents the analysis of high-resolution food chain for the mixed forest ecosystem of Khvalynsky National Park.

1. Introduction
Ecologists often want to understand what is the place of a species in an organic environment, its relation to food and enemies [1,2] or, more simply, Eltonian niche (hereinafter referred to as niche). Niches provide a semantic basis for linking species that inhabit the same environment. These species can be located along the hypothetical “niche axis”, which indicates the degree of their (species) similarity with each other. Species with overlapping niches compete for any resources associated with the niche axis, and therefore have a lower probability of harmonious coexistence [3]. In cases where one limiting resource can be used as a niche axis, this situation will become a simple basis for analyzing any ecological communities. However, in many cases, species require (and compete for) a wide range of abiotic and biotic resources that are not known to everyone. In such cases, it is almost impossible to identify niches for all species in the community. But despite this, it is possible to describe the biotic component of the niche of a species using food webs - the trophic interaction networks of the form [4]. These networks often describe antagonistic interactions, such as predation and parasitism, but may also include mutualisms (pollination, the spermatozoa), when one species feeds on another, while ensuring reproductive processes. Food networks describe energy and biomass flows through the community [5,6], show the ecosystem functions [7–9], and can give an idea of the overall situation of stability in the community [10, 11]. Thus, describing the role of species in food webs (i.e., how each species participates in its community) provides tools for assessing ecological niches of species, both in terms of their survival requirements and in terms of their impact on communities [12].

Various methods for analysis of food web have been actively used by researchers recently. One of the simplest mathematical definitions of the role of a species is its degree: the number of interaction partners (or food links, in which the species participate [13–23]). In addition to describing the ability of the species to influence the rest of the community, the degree can also
be used to give an idea of the vertical position of the species in the food web, i.e. its trophic level \([18,24–32]\). Some centrality indicators can be used to describe the ability of specific species to influence the rest of the food web \([14,33,33–37]\).

According to the data on the food chain of this location, we have identified the most important species, extinction of which will lead to the greatest violation of food web. To implement the task, we used graph theory which is discrete mathematical tools. Food chain represented in the form of an unweighted directed graph, where the vertices are the species, and directed edges are the couple who eats whom. We found the most important vertices by using the graph centrality measure algorithms: closeness centrality, betweenness centrality; and one of the reference ranking algorithms PageRank.

### 2. Measures of Centrality

One of the first ideas in the theory of network analysis is the idea of centrality (importance of the vertex in the graph). In the theory of analysis networks, the most important nodes are defined as central, and the measure of this centrality can be expressed numerically.

In other words, centrality is the importance of a node in a network, the characteristic of its position in relation to the other nodes and the network as a whole. This the structural importance of a node can help understand the cons and pros of its networking positions and also reflects how much the top participates in the information distribution process between the other vertices in the graph.

There are many indexes to assess the centrality of the vertices. Each of them is associated with a specific algorithm. In the context of this work we considered the following:

- Closeness centrality;
- Betweenness centrality.

#### 2.1. Closeness Centrality

Closeness centrality expresses how close a node is located to the rest of the network nodes. It is a measure of efficiency since the node that is closer to the other nodes of the graph is a better subject to the perception of new information and communication with the remaining vertices. In graph, closeness centrality (or closeness) of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. If some vertex is unreachable from this, distance to it is considered equal to the total number of vertices in the graph. Thus, the more central a node is, the closer it is to all other nodes. So closeness centrality \(C_C\) for the vertex \(x\) of the graph \(G = (V, E)\), where \(V\) is the set of vertices, \(|V| = n\), \(E\) is the set of edges, is calculated by formula:

\[
C_C(x) = \frac{1}{\sum_{y \in V} d(y, x)},
\]

where \(d(y, x)\) is the shortest distance from vertex \(y\) to \(x\). If they are not connected, then \(d(y, x) = n\).

Normalized centrality formulas are used to compare graphs with different numbers of vertices by multiplying \(C_C\) by \(n - 1\).

In an undirected graph, it does not matter to take a distance from this vertices to the rest \(d(x, y)\) or from others to a given \(d(y, x)\), because they are equal. While in directed graph it will lead to different results. Therefore two measures of closeness centrality for a directed graph are distinguished:

- on incoming links (IN \(d(y, x)\));
- by outgoing links (OUT \(d(x, y)\)).
In the framework of this task we used the option for incoming links, since the food chain is a uniquely directed system and the species most referenced by the others are prioritized over those who refer to many.

2.2. Betweenness Centrality

Betweenness centrality characterizes how important the node in the path between other nodes is. Betweenness centrality expresses how many shortest paths between all network nodes passes through a specific node. If some node has a high measure of this centrality that suggests it is the only connection between the various parts of the network.

For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

Betweenness centrality finds wide application in network theory: it represents the degree of which nodes stand between each other. For example, in a telecommunication network, a node with a higher betweenness centrality would have more control over the network, because more information will pass through that node. Betweenness centrality was devised as a general measure of centrality: it applies to a wide range of problems in network theory, including problems related to social networks, biology, transport and scientific cooperation.

For the graph \( G = (V, E) \), where \( V \) – the set of vertices (\(|V| = n\)), \( E \) is the set of edges, betweenness centrality \( C_B \) for the vertex \( v \in V \) is calculated by formula:

\[
C_B(v) = \sum_{\substack{s,t \in V \\ s \neq v \neq t}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}},
\]

where \( \sigma_{s,t} \) is the total number of shortest paths from node \( s \in V \) to \( t \in V \), \( \sigma_{s,t}(v) \) is the number of those paths that pass through \( v \).

Note that the betweenness centrality of a node scales with the number of pairs of nodes as implied by the summation indices. Therefore, the calculation may be rescaled by dividing through by the number of pairs of nodes not including \( v \), so that \( C_B(v) \in [0, 1] \). The division is done by \((n - 1)(n - 2)\) for directed graphs and \((n - 1)(n - 2)/2\) for undirected graphs. Note that this scales for the highest possible value, where one node is crossed by every single shortest path.

The measure of centrality in a graph is based on shortest paths and the assumption that if there are several shortest paths with the same length between two vertices each is used with equal probability.

3. PageRank

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of “measuring” its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references, thus to the whole graph.

The specificity of this parameter is that it is calculated depending on the weight of the reference mass. PageRank is total page authority, which depends on all sites that link to it. It is the weight of all links that determines such value as PageRank. It should be noted, that PageRank does not take into account the number of links, but their weight, so that a huge number of spammed links will not add value.

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be
calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called “iterations”, through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a “50% chance” of something happening. Hence, a PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

The PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor $d$. Various studies have tested different damping factors, but it is generally assumed that the damping factor will be set around $d = 0.85$. The damping factor is subtracted from 1 (and in some variations of the algorithm, the result is divided by the number of documents ($N$) in the collection) and this term is then added to the product of the damping factor and the sum of the incoming PageRank scores. The numerical weight that it assigns to any given element $E$ is referred to as the PageRank of $E$ and denoted by $PR(E)$.

In the aggregate we obtained the following equation:

$$PR(p_i) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)},$$

where $p_1, p_2, \ldots, p_n$ are the pages under consideration, $M(p_i)$ is the set of pages that link to $p_i$, $L(p_j)$ is the number of outbound links on page $p_j$, $N$ is the total number of pages, $d$ is damping factor.

To find the PageRank we used the iterative method of solving equations, since the record of the ranks of all pages can be represented in the form of a square system of linear algebraic equations.

At the start of the iterative process ($t = 0$) the initial approximation of ranks all pages are set to the same value:

$$PR(p_i; 0) = \frac{1}{N},$$

where $N$ is the total number of pages, $(p_i, 0)$ is page $p_i$ at the initial moment of time $t = 0$.

At each subsequent point in time ($t + 1$) PageRank is calculated by the ratio:

$$PR(p_i, t + 1) = \frac{1 - d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)}.$$  

The iteration process continues until the norm of the difference rank vectors of the current and previous approximation does not become insignificantly small.

4. Empirical Results
4.1. Input Data
Khvalynsky National Park has the status of a specially protected natural area of federal significance. The park was created on August 19, 1994 and is located in the Khvalynsky District, in the northeastern part of the Saratov Right bank. The natural boundary in the east is the river Volga, in the north and in the north-west the region borders with the Samara region, in the south with Volsky District. The total area of the park is 26,037 hectares.

To find the most important animal species, the data were obtained based on the food chain in the mixed forest Khvalynsky National Park (represented in Fig. 1), where each species is assigned its own unique number, and the list of what this species feeds on.
For further processing, we wrote all the data in a text file, numbering beforehand all species, so that the numbers started from zero. We implemented the necessary functionality with the help of the tools provided by Python programming language. Define a new class `myGraph` with fields:

- `Species` is a dictionary in which each number is associated with the name of the species;
- `G` is adjacency list of a directed graph, organized by a dictionary `sets`;
- `Q` is the set of names of all graph vertices.

The filling of these fields occurred by successively reading information from a text file. It is implemented in the initialization class method `def __init__(self, inputfile)`, where `inputfile` is the path to the document.

4.2. Visualization of the Food Chain
We have implemented a function `def draw(self, g)` that displays producers, consumers of the 1st, 2nd and 3rd orders and decomposers on their fixed niches (Fig. 2). We colored the vertices of each level differently for better visualization.

- producers in green;
- first order consumers in blue;
- second order consumers in orange;
- third order consumers in red;
- decomposers in white.

![Figure 2. View the full food chain](image)

Also, we changed the colors of the graph links:
- from first order consumer to green;
- from second order consumers to orange;
- from third order consumer to red;
- from decomposers to black.

The drawing functionality provided by the package igraph helped to realize this idea.

4.3. Closeness Centrality
To find the closeness centrality of the vertex $v$, we traversed all the vertices of the graph except $v$ and added the distance to the vertex $v$ to the resulting answer. This distance can be found using a breadth-first search algorithm since the graph is not weighted. After that, the total value is taken to be equal to the reciprocal of the found sum.

We found $C_C$ for all species and placed them in a list sorted in descending order (most significant at first) looking like this: $\{1, 2, 4, 5, 3, 0, 7, 12, 16, 11, 6, 17, 18, 19, 8, 10, 14, 9, 29, 28, 15, 27, 20, 21, 22, 30, 31, 32, 34, 35, 36, 13, 23, 24, 25, 26, 33, 37, 38, 39\}$. As expected, the producers (0–7) were at the top of this list, since the accumulation of energy begins with them and each animal has a path to them in the food chain. And animals that do not have natural enemies, such as decomposers (37–39), predators and large herbivores, were at the end of the list, because no one feeds on them, hence closeness centrality by incoming links is small for them.
Since we are more interested in the significance of animals in the food chain, but not plants, from the sorted list found the first three vertices which are consumers. They are: 12 (leaf-worm-fatty), 16 (pine sawfly), 11 (gypsy mole). We removed them from the food chain and saw how it would change, with the condition that no species would adapt. As we can see, the removal of these three species led to the disappearance of two second order consumers: 28 (spotted flycatcher) and 30 (common cuckoo), and also to the extinction of 31 links (Fig. 3).

4.4. Betweenness Centrality
To find betweenness centrality vertex \( v \) you need to know a list of all shortest paths between all pairs of vertices. For all possible pairs of vertices \( s, v \) we found the percentage of the paths passing through a given \( v \) to the total number of shortest paths and added this value to the final result, taking into account the fact that \( v \) is neither the initial \( s \) or final \( v \) vertices. We found betweenness centrality for all species and placed them in a list sorted in descending order (most significant at first) looking like this: \( \{12, 11, 17, 19, 15, 28, 29, 10, 18, 16, 14, 21, 22, 31, 36, 8, 30, 9, 32, 20, 0, 1, 2, 3, 4, 5, 6, 7, 13, 23, 24, 25, 26, 27, 33, 34, 35, 37, 38, 39\} \).

As expected, the species located at the periphery of the food chain, such as: producers (0–7), decomposers (37–39) and predators (32–36) have a low betweenness centrality, since in the calculation of the paths passing through a given one, the paths beginning or ending with it are not included. In turn, for these vertices almost all paths are as such. In the top of this list are the species belonging to the intermediate vertices of the food chain, that is, first order consumers. The first three species are: 12 (leaf-worm-fatty, Fig. 6), 11 (gypsy moth), 17 (yellow-necked wood mouse). By analogy with closeness centrality, we removed them from the food chain and looked at the changes. This led to the disappearance of one second order consumer 30 (common cuckoo) and 31 links (Fig. 4).
Figure 4. Betweenness centrality

4.5. PageRank
Finding the PageRank of each vertex of the graph is implemented using an iterative method for solving a system of linear equations. The damping factor is set to the theoretically recommended value $d = 0.85$. The initial approximation of the probability of meeting each vertex is set to the same value equal $1.0/n$, where $n$ is the total number of vertices. We begin the iterative process of finding a new approximation through the previous one. If the sum of the probabilities of the new approximation is not a unit, we make a linear normalization of all probabilities on the interval $[0, 1]$. We will continue the iteration process until the norm of the difference between the new and previous approximations becomes insignificantly small. At the end of the iterative process a list sorted in descending order (most significant at first) looking like this: {0, 1, 2, 4, 5, 7, 3, 12, 17, 18, 11, 16, 19, 8, 10, 14, 9, 29, 28, 6, 15, 27, 20, 21, 22, 30, 31, 32, 34, 35, 36, 13, 23, 24, 25, 26, 33, 37, 38, 39} was obtained.

As expected, the producers (0–7) are in the top of this list, since each animal species must have one of the primary sources of energy in the chain of its nutritional goals, which are producers. After them there are first and second order consumers. The list closes the list of species that do not have natural enemies, therefore the reference mass on them in the graph is insignificantly small, if not even zero. Such species are the highest predators (32–36), decomposers (37–39) and large herbivores. The first three species are: 12 (leaf-worm-fatty, Fig. 6), 17 (yellow-necked wood mouse), 18 (forest dormouse). By analogy with betweenness centrality, we removed them from the food chain and looked at the changes. This led to the disappearance of one second order consumer 27 (Nikolsky viper), and two third order consumers: 34 (tawny owl, Fig. 7), 35 (common buzzard); and 30 links (Fig. 5).
5. Conclusion
The most important species in the food chain were found along three measures of the centrality of the vertices in the graph:

- closeness centrality: 12 (leaf-worm-fatty), 16 (pine sawfly), 11 (gypsy mole);
- betweenness centrality: 12 (leaf-worm-fatty), 11 (gypsy moth), 17 (yellow-necked wood mouse);
- pagerank: 12 (leaf-worm-fatty), 17 (yellow-necked wood mouse), 18 (forest dormouse).

The partial intersection of the results of three different measures once again proves the correctness of the calculations performed to find the most important species in the mixed forest.
of the Khvalynsky National Park.

Leaf-worm-fatty and gypsy moth are forest pests characterized by ups and downs in their population size. Most forest insectivorous birds are polyphagous, i.e. have the ability to quickly switch to more accessible feed which are insect pests during population growth. Stereotypical reactions to its search and prey, produced by feeding with mass food, increase efficiency of the feeding chicks and are adopted by these species and other species of insectivorous birds. It leads to the concentration of consumer birds in areas of high numbers of insects and thereby increases the importance of insect-pests (Leaf-worm-fatty, gypsy moth) in food chain. However, this phenomenon is not observed every year. The rise of insect-pest population, as a rule, occurs under the influence of deviations of several meteorological indicators from the norm, most often within a few years.

Rodents (mice, voles) are extremely important in feeding predators. Many of the rodents are capable of producing rapid population growth. In the conditions of the mixed forest of the Khvalynsky National Park, the most common rodents are the yellow-necked wood mouse, the small forest mouse, the common vole, the forest dormouse. Small rodents are the main food for the absolute majority of forest predators. However, the majority of predators tend to be omnivorous, during the years of decline, the number of rodents switch to feeding on other feeds, such as frogs, lizards, small birds.

Acknowledgments
This work was supported by the Russian Fund for Basic Research, project 18-37-00060.

References
[1] Elton C 1927 Animal Ecology vol 365 (New York: Macmillan Co.)
[2] Johnson S D and Steiner K E 2000 Trends in Ecology & Evolution 15 140–143
[3] Godoy O, Bartomeus I, Rohr R P and Saavedra S 2018 Trends in Ecology & Evolution 33 287–300
[4] Pinam S L, Lawton J H and Cohen J E 1991 Nature 350 669–674
[5] Lindeman R L 1942 Ecology 23 399–417
[6] Wootton J T 1997 Ecological Monographs 67 45–64
[7] Memmott J, Craze P G, Waser N M and Price M V 2007 Ecology Letters 10 710–717
[8] Reiss J, Bridle J R, Montoya J M and Woodward G 2009 Trends in Ecology & Evolution 24 505–514
[9] Thompson R M, Brose U, Dunne J A, Hall R O, Hladysz S, Kitching R L, Martinez N D, Rantala H, Romanuk T N, Stouffer D B and Tylianakis J M 2012 Trends in Ecology & Evolution 27 689–697
[10] Neutel A M 2002 Science 296 1120–1123
[11] Thebault E and Fontaine C 2010 Science 329 853–856
[12] Chase J M and Leibold M A 2003 Ecological Niches (Chicago: University of Chicago Press)
[13] Jordan F, Benedek Z and Podani J 2007 Ecological Modelling 205 270–275
[14] Lai S M, Liu W C and Jordán F 2012 Biology Letters 8 570–573
[15] Curtsdotter A, Binzer A, Brose U, de Castro F, Ebenman B, Eklof A, Riede J O, Thierry A and Rall B C 2011 Basic and Applied Ecology 12 571–580
[16] Dunne J A, Williams R J and Martinez N D 2002 Ecology Letters 5 558–567
[17] Memmott J, Waser N M and Price M V 2004 Proceedings of the Royal Society of London B: Biological Sciences 271 2605–2611
[18] Eklof A and Ebenman B 2006 Journal of Animal Ecology 75 239–246
[19] Kaiser-Bunbury C N, Muff S, Memmott J, Miller C B and Caflisch A 2010 Ecology Letters 13 442–452
[20] Allesina S 2012 Nature 487 175–176
[21] Blüthgen N, Menzel F, Hovestadt T and Fiala B 2007 Current Biology 17 341–346
[22] Dormann C F 2011 Network Biology 1 1–20
[23] Nilsson K A and McCann K S 2016 Theoretical Ecology 9 59–71
[24] Thompson R M, Hemberg M, Starzomski B M and Shurin J B 2007 Ecology 88 612–617
[25] Williams R J and Martinez N D 2000 Nature 404 180–183
[26] Lindeman R L 1942 Ecology 23 399–417
[27] Darnell R M 1961 Ecology 42 553–568
[28] Baird D and Ulanowicz R E 1989 Ecological Monographs 59 329–364
[29] Christian R R and Luczkovich J J 1999 Ecological Modelling 117 99–124
[30] Dyer L A and Letourneau D 2002 Ecology Letters 6 60–68
[31] Boersma K S, Bogan M T, Henrichs B A and Lytle D A 2014 Oikos 123 807–816
[32] Estes J A, Burdin A and Doak D F 2015 Proceedings of the National Academy of Sciences 113 880–885
[33] Estrada E 2007 Journal of Theoretical Biology 244 296–307
[34] Jordan F, Liu W c and Davis A J 2006 Oikos 112 535–546
[35] Plank M J and Law R 2011 Theoretical Ecology 5 465–480
[36] Donohue I, Petchey O L, Montoya J M, Jackson A L, McNally L, Viana M, Healy K, Lurgi M, O’Connor N E and Emmerson M C 2013 Ecology Letters 16 421–429
[37] Rohr R P, Saavedra S and Bascompte J 2014 Science 345 1253497–1253497