On Building Online Visualization Maps for News Data Streams by Means of Mathematical Optimization

Emilio Carrizosa (PhD)[1], Vanesa Guerrero (PhD) [† 2], Daniel Hardt (PhD)[3], and Dolores Romero Morales (PhD)[3]

[1]Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Seville, Spain

ecarrizosa@us.es

[2]Departamento de Estadística, Universidad Carlos III de Madrid, Getafe, Spain

vanesa.guerrero@uc3m.es

[3]Copenhagen Business School, Frederiksberg, Denmark

{dh.digi, drm.eco}@cbs.dk

[†]Corresponding Author

Abstract

In this paper we develop a novel online framework to visualize news data over a time horizon. First, we perform a Natural Language Processing analysis, where the words are extracted, and their attributes, namely the importance and the relatedness, are calculated. Second, we present a Mathematical Optimization model for the visualization problem and a numerical optimization approach. The model represents the words using circles, the time-varying area of which displays the importance of the words in each time period. Word location in the visualization region is guided by three criteria, namely, the accurate representation of semantic relatedness, the spread of the words in the visualization region to improve the quality of the visualization, and the visual stability over the time horizon. Our approach is flexible, allowing the user to interact with the display, as well as incremental and scalable. We show results for three case studies using data from Danish news sources.

Keywords: Online Visualization; News Data Streams; Word Importance; Relatedness; Visual Stability

1 Introduction

With the growth of Big Data, new business opportunities arise for gaining insight from data streams [2, 21, 29, 55], and Information Visualization has played a crucial role in facilitating this task, [12, 15, 26, 27, 42, 59]. However, this can be very challenging, not just because of the size of these data streams [1, 6, 31, 50], but also because of structural complexities that make the analysis difficult [8, 9, 22, 34]. This is the case for news data [16, 19, 23, 35, 36, 44, 45, 49, 57], where there is a virtually unlimited supply of data available, containing insights that are crucial for business and society [13]. In this paper, we show how these insights can be unlocked from news data by addressing its structural complexity.

News data can be considered as a sequence of words, which can be visualized using methods such as word clouds. Word clouds aim to represent the frequency or importance of words in texts by balancing several aesthetic criteria [61, 62], including avoiding overlapping and reducing the empty spaces. However, standard word clouds do not manage to convey two important features of the structure of news data. The first is **temporal** – it is important to know how news topics develop over time. The second is **semantic** – certain words and topics are closely related to other words or topics.

In this paper we use Mathematical Optimization techniques [7, 10] to build a novel online visualization map that makes it possible to produce visualizations of news data as it develops over time. This online visualization in turn depends on Natural Language Processing for uncovering the semantic structure in the news data, by computing the importance and the relatedness of the words that occur. In this way a kind of dynamic word cloud is produced, which displays the importance as well as the relatedness of words, and shows how they develop over time. Although there have been attempts in the literature to incorporate the temporal [20] and the semantic [4, 24, 66] features into word clouds, as far as the authors are aware, this is the first paper in which these are optimized

simultaneously. The approach proposed here has important advantages. First, by being online, and therefore incremental, we do not require the recalculation of the past display whenever new data comes in, [36]. Second, our approach is scalable since the optimization problems solved in each iteration are computationally cheap. Third, the modeling is flexible, allowing the user to interact with the display, for example, by defining new criteria to be optimized, constraints on the layout of the display, and different visualization regions over time. Finally, our work employs word embeddings [41, 51, 52, 53, 56], a newly developed method that provides a much more nuanced and precise model of semantic structures.

We consider a set $\mathcal{N}$ of words and T time periods. Our aim is to build an online visualization map, i.e., a collection of T visualization maps, where the words in $\mathcal{N}$ are represented in each time period. We have two attributes associated with the words, the so-called importance and the relatedness matrix, which are both dynamic. For each word i and each period t, $\omega_{i,t}$ represents the importance of word i in time period t. For each pair of words i and j and each period t, $\delta_{ij,t}$ represents the semantic relatedness between words i and j in period t. In each period, we depict words as circles representing accurately the importance $\omega_{i,t}$. The position of the circles is defined by three criteria. The first criterion takes care that, in each time period, the distance between the circles resembles the relatedness $\delta_{ij,t}$ attached to the words, as in MultiDimensional Scaling [37, 38, 60]. The second criterion aims at spreading the circles across the visualization region to improve the aesthetics [62] and ensure the readability of the online visualization map. The third criterion ensures a smooth transition between the visualization maps, thus ensuring visual stability, [18]. Finally, the approach is online, such that the plots for the first $T_0 - 1$ periods are already available, and we need to construct the plots in periods $T_0, \ldots, T$, where new words may appear, using the criteria above. With these criteria, our ordered time sequence of visualizations behaves as a kind of animated conceptual space, helping to show how words gain or recede in importance

over time.

The remainder of the paper is structured as follows. In Section 2, we review the related literature. In Section 3, we describe the way the importance and the relatedness are calculated in this paper. In Section 4, we describe the Mathematical Optimization approach used to build the online visualization map. In Section 5, we illustrate the usefulness of our tool, by applying it to three case studies from Danish news. Some concluding remarks and plans for future research are given in Section 6.

2 Literature review

In this paper, we describe an online visualization framework based on Mathematical Optimization, which can extract insights from news data streams as they develop over time. Mathematical Optimization is typically used to represent dissimilarities (the complement of relatedness), see e.g., [5, 28, 46, 47] and references therein. However, we are not aware of any similar approaches to extracting these insights from news data streams.

The framework involves several key elements:

- Importance: term importance must be computed, both to decide which terms to include, and their relative size
- Relatedness: the semantic relatedness of terms must be calculated – this provides a basis for deciding where to place terms in the visualization region
- Visual Stability: this is made possible by the computation of semantic relatedness, together with a Mathematical Optimization model which ensures that related terms appear close to each other. As a side effect of this, term position will remain stable in a temporal sequence of visualizations. This stability over time is crucial to providing the insights over time from

a news data stream.

In this section, we will review related work which addresses certain key elements of this approach, although none of them put them together in the way described here.

Word Clouds have become a widespread tool for visualizing the important terms or concepts in texts. *Wordle* [33, 62] is a widely used tool for this purpose. There have been many works to improve word clouds as data exploratory tool, see [24] and references therein. In [64], *EdWordle* is proposed to allow the user to edit the word cloud (e.g., adding/deleting, dragging, resizing words), while preserving its neighborhood structure. After editing, the words are rearranged based on rigid body dynamics, [65]. In [14], this consistency of the neighborhood structure is applied to dynamic data, where the shape defining the visualization region of the cloud may vary over time too. In [11], the authors develop a methodology to coordinate word clouds. They introduce the concept of a *word storm*, i.e., a collection of word clouds, each of them associated with a different text. Since word storms are used to compare documents, the authors argue that the same word should have a similar location across different word clouds. One of the approaches to build word storms is based on Mathematical Optimization. The objective function has three goals, namely, texts that are similar are represented by similar word clouds, the frequencies are represented accurately, and the aesthetics of the display (measured by compactness and avoidance of overlapping).

Normally word clouds provide information along one dimension, namely, the frequency or importance of the term, which is represented by the size of the term in the display. Most of the effort is put into the aesthetics of the word cloud. The position of a term in the display is essentially random; it conveys no information. In the present approach, the position of the term conveys important information – namely, semantically related terms appear closer to each other, and unrelated terms are more distant.

Some recent work describes word clouds that attempt to incorporate semantic relatedness. For

example, Cui et al. [20] use semantic relatedness to visualize temporal document content evolution. The authors propose building a dynamic word cloud and a trend chart to highlight *significant* changes in content over time. They use a stepwise procedure to build the dynamic word cloud. First, MultiDimensional Scaling (MDS) is executed on the set of all words across all periods to choose the initial points where the words will be anchored. The authors define three possible dissimilarities to be used by MDS, one of them being based on semantic relatedness using feature vectors as in [58]. Second, a word cloud is built for each period, starting with the points corresponding to words appearing in that period, and, subsequently, the aesthetics of the display is improved with the so-called force-directed algorithm. As the authors themselves note, their approach does not allow for user interaction.

Barth et al. [4] propose three approaches for building semantics-preserving word clouds, i.e., word clouds trying to represent accurately both word frequencies as well as relatedness between words. The first one starts with an MDS solution built on the rescaled problem, where both the frequencies and the dissimilarities are scaled by the same constant. The word cloud is inflated iteratively and the overlapping is removed using the procedure in [20]. The two other approaches are built around the graph defined by the dissimilarities. The second one builds a collection of stars, i.e., trees of depth one. Then, an MDS layout is built for the stars, while the force-directed algorithm is called to improve the aesthetics of the display as above. The third approach works similarly, but in this case the authors extract cycles/paths from the dissimilarity graph. Barth et al. [3] is a follow-up work to [4]. The authors provide a graph representation of the problem, where the vertices are the words and the weights on the arcs are the relatedness between them, they then show that the problem is NP-Complete, and propose approximation algorithms for specific structures.

Further approaches in terms of text visualizations and analysis are surveyed in [43].

3 Natural Language Processing analysis

Our aim is to produce an online visualization map, which provides semantic insights into a stream of news data. This requires Natural Language Processing analysis from two perspectives: **Importance** and **Relatedness**.

3.1 Importance

Importance involves extracting the most interesting words appearing in the news data stream, and providing a numerical rating of the importance of the word. In other words, we want to know whether the word indicates an important topic in the current news data stream.

Computing semantic importance of words is a well-studied problem, and we follow well-established techniques, with some modifications to address issues related to our application, in which we wish to extract insights from a topical news data stream.

The frequency of word occurrences provides a basic, yet flawed, measure of semantic importance. An obvious flaw concerns extremely common words that typically express particular linguistic functions, such as prepositions (in, on, of), articles (a, an, the) or conjunctions (and, or). Despite their frequency, such words do not indicate semantically important topics. It is typical to define a *stopword list* consisting of such words.

The problem of stopwords indicates a more general problem – frequency simply does not correlate with semantic importance. A standard way of addressing this is to use a different metric: Term Frequency Inverse Document Frequency ($tfidf$). The idea of this metric is to reward words for appearing frequently in the particular text of interest, but to penalize words that generally tend to appear frequently across many texts.

This penalty is captured by the inverse document frequency (idf) [48], which is

$idf = log(D/d)$

where D is the total number of documents, and d is the number of documents where the term occurred. Thus, idf is smaller for terms that occur in many documents. The $tfidf$ value is computed with the following formula

$$tfidf = tf * idf$$

where tf is simply the number of times the term appears in the current document – in our case the selected news data stream. In the current study, we have selected several topical news data streams, by selecting all articles in a certain period mentioning a certain key term, such as `Internet`, `Terror Attack` or `Immigration`. So in these cases, the document is considered to be the currently selected news stream, and we compute term frequency as all the occurrences of a term within that news data stream. To compute the inverse document frequency, idf, we selected a large collection of proceedings from the European Parliament (Europarl) [32]. The Europarl collection is a standard resource in Natural Language Processing, because it provides a large reference for standard language use for all the major European languages. It is thus excellent for our purpose, which is to measure the typical frequency of words across many texts.

The importance of a given term i, for a given time period t, is given by:

$$\omega_{i,t} = tf_{i,t} * idf_i$$

where $tf_{i,t}$ is computed for the news text of interest in time period t, and idf_i is computed in the general background text, namely the European Parliament proceedings.

3.2 Relatedness

It is not enough to compute the relative importance of terms in a news stream. Some terms are closely related to other terms. For example, in the news data stream on the keyword `Immigration`, the terms *muslim*, *scarf*, and *immigrant background* are closely related.

It is widely acknowledged that cooccurence data can provide some valuable information about

the meaning of a term, a kind of semantic signature, and that similarity of the context of two words provides a guide to their degree of semantic relatedness [25]. We compute the relatedness between terms by using *word embeddings* for each term, i.e., high-dimensional vectors, using the Python `gensim` package [56], which implements *word2vec*. Using *word2vec*, the word embedding is built by training a neural network to predict the probability of cooccurence for pairs of words – given a pair of words such as, say, *table* and *chair*, the network yields the probability of these two words in fact coocurring. By cooccurence, we mean that the two words appear either adjacent to each other, or separated by at most ℓ words, where we have chosen $\ell = 3$ in our case studies in Section 5. The neural network is constructed through a training phase, in which the network is presented with a large corpus of text data. In this case, we used a corpus of data consisting of 55,000 news articles from the online site of Jyllands-Posten, a major Danish newspaper.

The training process of the network involves setting weights for a large number of features for each word in the vocabulary – in our case studies, this number is set to 300. This results in a sequence of 300 real-valued numbers for each word in the vocabulary. We can then compare these values for different words – since the values have been set to predict cooccurence observations, one might expect that similar words would tend to have similar values. Indeed, this expectation has been confirmed in a great deal of recent work, see, for example, [40, 54].

We use the `gensim` method, `similarity`, which computes the cosine distance between two vectors. This gives a value ranging from -1 to +1, with higher values denoting higher degrees of similarity. We then convert this to a *dissimilarity* value ranging between 0 and 1, with 0 the most similar and 1, the most different. Thus for a given pair of terms i and j and a given time period, we define the dissimilarity $\delta_{ij,t}$ between terms i and j in period t as follows:

$$\delta_{ij,t} = 1 - (similarity_{ij} + 1)/2,$$

where $similarity_{ij}$ is the value returned by `similarity` when applied to terms i and j. Note that

this dissimilarity is time-independent, but our approach would still be valid for time-dependent ones.

The dissimilarity, and in turn the relatedness, provides a crucial element to our visualization model, namely a basis for grouping terms together, or placing them far apart. Such groupings can provide insightful generalizations about categories of terms that occur in a given news data stream. As described below, we produce a visualization where terms are placed on the visualization region in a way that is as faithful as possible to their semantic relatedness, while a given term will tend to appear in the same area on the visualization region. Because of this, an ordered time sequence of visualizations will behave as a kind of animated conceptual space, helping to show how terms gain or recede in importance over time.

4 The visualization model

In this section, we present a Mathematical Optimization model and a numerical optimization approach for the visualization in the region $\Omega = [0,1] \times [0,1]$ of the words in $\mathcal{N}$ and the attributes returned by the Natural Language Processing analysis $\omega_{i,t}$ and $\delta_{ij,t}$. For convenience, we denote the cardinality of $\mathcal{N}$ by N. Ours is an online approach: the first $T_0 - 1$ plots are already available for words appearing in those periods, and our goal is to construct the next $T - (T_0 - 1)$ plots, for words appearing in those periods, $\mathcal{N}$. Without loss of generality, we assume that the first N_0 words in $\mathcal{N}$ appear in the time horizon $\{T_0 - S, \ldots, T_0 - 1\}$, with $S \leq T_0 - 1$. These N_0 words will play an important role when ensuring visual stability with the plots at hand.

We display the words using circles. Let τ be a common positive scale factor for all circles and all periods. For each i and t, word i is represented in the visualization map associated with time period t by the circle $\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t})$ centered at $\boldsymbol{c}_{i,t} \in \mathbb{R}^2$ and of radius $\tau\omega_{i,t}$. This means that the importance $\omega_{i,t}$ is represented by the area of circle $\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t})$ exactly up to the scale factor τ, which

can be seen as an aesthetics parameter to be chosen by the user to control the amount of area of Ω to be covered by the circles. The dissimilarity $\delta_{ij,t}$ is represented by the infimum distance between circles $\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t})$ and $\mathcal{C}_{j,t}(\boldsymbol{c}_{j,t})$, up to the scale factor κ, defined as

$$d(\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t}), \mathcal{C}_{j,t}(\boldsymbol{c}_{j,t})) = \min_{\boldsymbol{b}_{i,t} \in \mathcal{C}_{i,t}(\boldsymbol{c}_{i,t}), \boldsymbol{b}_{j,t} \in \mathcal{C}_{j,t}(\boldsymbol{c}_{j,t})} \|\boldsymbol{b}_{i,t} - \boldsymbol{b}_{j,t}\|,$$

where $\|\cdot\|$ denotes the Euclidean norm.

The location of the circles in the visualization region is guided by the weighted average of three criteria, $F = \lambda_1 F_1 + \lambda_2 F_2 + \lambda_3 F_3$, $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. In turn, they model the accurate representation of the dissimilarities, the spread of the words in the visualization region to improve the aesthetics of the visualization, as well as the visual stability over the time horizon. In what follows, we formalize the definition of the three criteria, see also [7, 10].

To quantify the resemblance of the distances between circles in the online visualization map to the dissimilarities between the corresponding words, we use the summation of squared errors, yielding

$$F_1(\boldsymbol{c}_{1,T_0}, \ldots, \boldsymbol{c}_{N,T}) = \sum_{t=T_0}^{T} \sum_{\substack{i,j \in \mathcal{N}(t) \\ i \neq j}} \left[d(\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t}), \mathcal{C}_{j,t}(\boldsymbol{c}_{j,t})) - \kappa \delta_{ij,t} \right]^2,$$

where $\mathcal{N}(t)$ is the set of words appearing in period t, $t = T_0, \ldots, T$.

To ensure the spread of the circles across the visualization region Ω, we use the summation of the squared distances, yielding in a minimization form

$$F_2(\boldsymbol{c}_{1,T_0}, \ldots, \boldsymbol{c}_{N,T}) = -\sum_{t=T_0}^{T} \sum_{\substack{i,j \in \mathcal{N}(t) \\ i \neq j}} d^2(\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t}), \mathcal{C}_{j,t}(\boldsymbol{c}_{j,t})).$$

To ensure a smooth transition between close visualization maps, we want words appearing in a similar location across the time. We model this through the distance between centers. We denote by S the smoothing window and by $\eta_s \geq 0$, $s = 1, \ldots, S$, the smoothing parameters. We use the

distance between $\boldsymbol{c}_{i,t}$ and the ones in the S previous periods, $\boldsymbol{c}_{i,t-s}$, scaled by η_s, yielding

$$\begin{aligned} F_3(\boldsymbol{c}_{1,T_0-S},\ldots,\boldsymbol{c}_{N_0,T_0-1},\boldsymbol{c}_{1,T_0},\ldots,\boldsymbol{c}_{N,T}) \;=\; & \sum_{t=T_0}^{T}\sum_{s=1}^{S}\eta_s \sum_{i=1,\ldots,N_0} \|\boldsymbol{c}_{i,t}-\boldsymbol{c}_{i,t-s}\|^2 \\ & + \sum_{t=T_0}^{T}\sum_{s=1}^{\min\{S,t-T_0\}}\eta_s \sum_{i=N_0+1,\ldots,N} \|\boldsymbol{c}_{i,t}-\boldsymbol{c}_{i,t-s}\|^2. \end{aligned}$$

In summary, our Mathematical Optimization model seeks the values of the variables $\boldsymbol{c}_{i,t}$, $i = 1\ldots N; t = T_0,\ldots,T$, so that circles $\mathcal{C}_{i,t}(\boldsymbol{c}_{i,t})$ fall inside Ω, and the F is minimized. The Online Visualization Map (OnViMap) problem is stated as follows

$$\begin{aligned} \text{minimize} \quad & F(\boldsymbol{c}_{1,T_0-S},\ldots,\boldsymbol{c}_{N_0,T_0-1},\boldsymbol{c}_{1,T_0},\ldots,\boldsymbol{c}_{N,T}) \\ \text{s.t.} \quad & \mathcal{C}_{i,t}(\boldsymbol{c}_{i,t}) \subseteq \Omega,\ i=1,\ldots,N; t=T_0,\ldots,T \\ & \boldsymbol{c}_{i,t} = \bar{\boldsymbol{c}}_{i,t},\ \ i=1,\ldots,N_0; t=T_0-S,\ldots,T_0-1 \\ & \boldsymbol{c}_{i,t} \in \mathbb{R}^2,\ i=1,\ldots,N; t=T_0,\ldots,T, \end{aligned} \qquad (OnViMap)$$

where $\bar{\boldsymbol{c}}_{i,t}$, $i = 1,\ldots,N_0; T_0-S,\ldots,T_0-1$, are centers in the plots at hand for the first T_0-1 periods.

Our numerical approach has this Mathematical Optimization model as the basis. Using the results in [7], we can show that (OnViMap) has a difference of convex objective function with an amenable decomposition, namely a separable function with quadratic and linear terms. In addition, the feasible region is defined by box constraints. Both of these features ensure an efficient implementation of the Difference of Convex Algorithm [39], an iterative procedure in which the concave term in the objective function is replaced by a linear majorization, yielding problems that are very tractable, namely convex quadratic problems in one variable.

Therefore, our approach is flexible, since it is based on Mathematical Optimization modeling, allowing for user interaction, being able to incorporate, e.g., constraints on the position of certain words. It is also incremental, since new words can be incorporated as they arrive, and it is scalable, since the problems solved in each iteration are computationally cheap. In addition, our approach

can easily incorporate other desirable features. It is possible to model other criteria to guide the optimization, such as the compactness of the display and the avoidance of overlapping; to use other types of glyphs to represent the words, such as rectangles, whose size would depend on the importance of the words but also on their lengths, [7]; and to have dynamic visualization regions that change over time, [14].

5 Case studies

5.1 Introduction

We have described a tool which displays the importance of terms as well as the dissimilarities between them, and produces visualizations based on news streams that develop over time. We have applied the tool to three case studies using data from Danish news sources. In Section 5.2, we explain how the data for the visualization has been extracted and how the parameters of the mathematical modeling have been set. In Section 5.3, we present the online visualization map for each case.

5.2 Data and parameter setting

Each case study is defined by a Danish keyword and a time frame. We have chosen *Internet* (also `Internet`, in English) from 1994-1997, *Terrorangreb* (`Terror Attack`, in English), from 2001-2015, and *Indvandring* (`Immigration`, in English), from 1995-2015. The news data streams are collected using the Infomedia Media Archive [30], a comprehensive collection of media sources in Denmark.

For each keyword, we select all articles in each period of the given time frame that contain the keyword, and extract the words from those articles. For *Internet*, this yields 7,033 articles, containing 4,457,420 words; for *Terrogangreb*, 27,549 articles, containing 20,276,933 words; and for *Indvandring*, 12,286 articles, containing 10,994,755 words.

The importance of each word is calculated as described in Section 3.1. The set of words to be plotted, $\mathcal{N}$, consists of the 20 most important ones in each period. For *Internet*, we have $N = |\mathcal{N}| = 40$; for *Terrogangreb*, $N = 127$; and for *Indvandring*, $N = 202$. The dissimilarity between the words in $\mathcal{N}$ is calculated as described in Section 3.2. The words, their importance and dissimilarity can be found in https://dataverse.harvard.edu\/privateurl.xhtml?token=63d03f6c-c444-4e27-a1e3-3c242cb1bc06.

In the Mathematical Optimization modelling, we have chosen as scale factors

$$\tau = \frac{1}{\max_t \sum_{i=1}^{N} \omega_{i,t}} \cdot 0.10$$

$$\kappa = \frac{N(N-1)T}{\sum_{t=1}^{T} \sum_{\substack{i,j=1,\dots,N \\ i \neq j}} \delta_{ij,t}} \cdot 0.30.$$

To ensure visual stability, we have chosen $S = 1$ and $\eta_1 = 1$. In the objective function F, we have chosen $\lambda_1 = 0.45, \lambda_2 = 0.15, \lambda_3 = 0.40$. Finally, and without loss of generality, we assume that $T_0 = 1$, i.e., there are no plots at hand from previous periods.

5.3 The visualizations

In this section, we present the visualization of each case study. Note that the label corresponding to each word has a font size proportional to its importance.

For the Danish keyword *Internet* (`Internet`), there are $N = 40$ words, which can be found in Table 1 as well as their translation into English. To give an idea of how these words cluster around each other using the dissimilarities, a hierarchical Cluster Analysis has been performed, yielding the dendrogram depicted in Figure 1. The visualization map can be found in Figures 2-5, where words appearing new with respect to the previous period are underlined, to help the user with this identification process. The importance of the words is clearly time-dependent, and there are some

words that only feature in one period out of the four periods, such as 'Diatel' (1995), a Danish videotex online service accessible through telephone lines, and 'Telia' (1996), a telephone company in Denmark. Even though the dissimilarity between words is time-independent, it is affected by words entering/leaving set $\mathcal{N}(t)$ from one period to the next. Nevertheless, our plots preserve these dissimilarities reasonably well. From the dendrogram in Figure 1, we can see words 'Telia', 'IBM', 'Apple', 'Microsoft', 'Diatel', 'PBS', 'Tele_Danmark' and 'UNI-C' clustering together. In the visualization map, these words are placed on the right side of the plot and close to each other. This is especially true for 'Apple' and 'Microsoft', which are present in all 4 periods.

For the Danish keyword *Terrorangreb* (`Terror Attack`), Table 2 contains the $N = 127$ words and their translation into English, while their dendrogram can be found in Figure 6. The visualization map can be found in Figures 7-21. The words 'terrorangreb' and 'Al-Qaeda' feature prominently in each period, and they are located close to each other on the left-top corner of the visualization region. Examples of words that feature in fewer periods are 'tegninger' (drawings) and 'profeten' (the Prophet), that appear for the first time in 2005 and reappear in 2015. From the dendrogram in Figure 6, we can see words 'PET', 'PST', 'Pentagon', 'FBI', 'CIA' and 'NSA' clustering together. In the visualization map, these words are placed around the anti-diagonal of the visualization region and close to each other. This is especially true for 'PET' (Danish Security and Intelligence Service) and 'CIA', which are present in almost all periods.

For the Danish keyword *Indvandring* (`Immigration`), Table 3 contains the $N = 201$ words and their translation into English, while their dendrogram can be found in Figure 22. The visualization map can be found in Figures 23-43. The keyword `Immigration` shares several words with the previous one (`Terror Attack`), including the names of several Danish and Swedish political parties. The political parties are clustered around each other in the dendrogram, 'Ny_Alliance', 'Sverigedemokraterne', 'Kristeligt_Folkeparti', 'Fremskridtspartiet', 'Dansk_Folkeparti' (also known

as DF), 'SF' and 'Venstre'. In the visualization map, these parties appear close to each other on the right hand side of the visualization region.

6 Conclusions

In this paper we have developed a novel and flexible online framework for the visualization of news data streams. We have combined techniques from Natural Language Processing and Mathematical Optimization to extract the most relevant words associated with a given keyword, compute their importance and their relatedness, and build an online visualization map. In this map, the words have been represented by circles, the area of which indicate the importance of words accurately, and that are placed in the visualization region following three criteria, namely, resemblance of the distances between circles to the dissimilarities, spread of the circles across the visualization region and stability of the position of the circles across the time horizon. Our Mathematical Optimization approach can incorporate constraints defined by the user, other criteria to guide the optimization, other glyphs to represent the words, and a dynamic visualization region that changes over time. We have presented three case studies, using data from Danish news sources, and based on these studies we argue that the tool provides a level of insight beyond that of previously described work.

There are several interesting lines for future research. First, the modeling approach in this paper is flexible and allows for desirable functionalities of data exploratory tools, such as the comparison of topics [17]. This boils down to modeling visual stability across topics, similar to the way in which we have modelled the visual stability across time, ensuring that words appear in a similar location across different topics. Second, the approach proposed in this paper can be extended to cope with other types of data streams that exhibit both temporal and semantic structure. For example, product reviews are similar to news data in possessing temporal and semantic structure. The techniques described in this paper could therefore be useful in *review mining* to gain insights in

customer experience [63]. Third, we plan to explore more sophisticated linguistic structures in our analysis, which currently is restricted to individual, frequently occurring words or short phrases. In future work we will provide a more systematic treatment of multiword phrases. Fourth, in the current work the importance of a given word is based solely on the occurrences of that word. A more sophisticated treatment might involve clustering of similar words, so that what is displayed is a label selected to represent the importance of a group of similar words.

Acknowledgement. *This research is funded in part by Projects MTM2015-65915-R (Spain), P11-FQM-7603 and FQM-329 (Andalucía), all with EU ERD Funds.*

References

[1] C. Albrecht-Buehler, B. Watson, and D.A. Shamma. Visualizing live text streams using motion and temporal pooling. *IEEE Computer Graphics and Applications*, 25(3):52–59, 2005.

[2] B. Baesens. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley and SAS Business Series. Wiley, 2014.

[3] L. Barth, S.I. Fabrikant, S.G. Kobourov, A. Lubiw, M. Nllenburg, Y. Okamoto, S. Pupyrev, C. Squarcella, T. Ueckerdt, and A. Wolff. Semantic word cloud representations: Hardness and approximation algorithms. In A. Pardo and A. Viola, editors, *LATIN 2014: Theoretical Informatics*, volume 8392 of *Lecture Notes in Computer Science*, pages 514–525. Springer Berlin Heidelberg, 2014.

[4] L. Barth, S.G. Kobourov, and S. Pupyrev. Experimental comparison of semantic word clouds. In J. Gudmundsson and J. Katajainen, editors, *Experimental Algorithms*, volume 8504 of *Lecture Notes in Computer Science*, pages 247–258. Springer International Publishing, 2014.

[5] A. Boytsov, F. Fouquet, T. Hartmann, and Y. LeTraon. Visualizing and Exploring Dynamic High-Dimensional Datasets with LION-tSNE. *arXiv preprint arXiv:1708.04983*, 2017.

[6] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2649–2658, 2012.

[7] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualization of complex dynamic datasets by means of mathematical optimization. Technical report, IMUS, Sevilla, Spain, 2017.

[8] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing proportions and dissimilarities by space-filling maps: a large neighborhood search approach. *Computers & Operations Research*, 78:369–380, 2017.

[9] E. Carrizosa, V. Guerrero, and D. Romero Morales. On mathematical optimization for the visualization of frequencies and adjacencies as rectangular maps. *European Journal of Operational Research*, 265:290–302, 2018.

[10] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming*, 169:119–140, 2018.

[11] Q. Castellà and C. Sutton. Word storms: Multiples of word clouds for visual comparison of documents. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14, pages 665–676. ACM, 2014.

[12] C.P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

[13] H. Chen, R.H.L. Chiang, and V.C. Storey. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165–1188, 2012.

[14] M.-T. Chi, S.-S. Lin, S.-Y. Chen, C.-H. Lin, and T.-Y. Lee. Morphable word clouds for time-varying text data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(12):1415–1426, 2015.

[15] J. Choo and H. Park. Customizing computational methods for visual analytics with big data. *IEEE Computer Graphics and Applications*, 33(4):22–28, 2013.

[16] C. Collins, F.B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009*, pages 91–98, 2009.

[17] G. Coppersmith and E. Kelly. Dynamic wordclouds and vennclouds for exploratory data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 22–29, 2014.

[18] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290, 2014.

[19] W. Cui, H. Qu, H. Zhou, W. Zhang, and S. Skiena. Watch the story unfold with textwheel: Visualization of large-scale news streams. *ACM Transactions on Intelligent Systems and Technology*, 3(2):1–17, 2012.

[20] W. Cui, Y. Wu, S. Liu, F. Wei, M.X. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53, 2010.

[21] T.H. Davenport. *Big Data @ Work*. Harvard Business Review Press, 2014.

[22] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[23] T. Gao, J.R. Hullman, E. Adar, B. Hecht, and N. Diakopoulos. NewsViews: an automated

pipeline for creating custom geovisualizations for news. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3005–3014. ACM, 2014.

[24] E. Gomez-Nieto, W. Casaca, D. Motta, I. Hartmann, G. Taubin, and L.G. Nonato. Dealing with multiple requirements in geometric arrangements. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1223–1235, 2016.

[25] Z.S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[26] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53:59–67, 2010.

[27] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30–55, 2012.

[28] G.E. Hinton and S.T. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, pages 857–864, 2003.

[29] A.H. Huang, R. Lehavy, A.Y. Zang, and R. Zheng. Analyst information discovery and interpretation roles: A topic modeling approach. Forthcoming in *Management Science*, 2017.

[30] Infomedia. www.infomedia.dk, 2017.

[31] A. Jakaitiene, M. Sangiovanni, M.R. Guarracino, and P.M. Pardalos. *Multidimensional Scaling for Genomic Data*, pages 129–139. Springer International Publishing, Cham, 2016.

[32] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86, 2005.

[33] K. Koh, B. Lee, B. Kim, and J. Seo. ManiWordle: Providing Flexible Control over Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1190–1197, 2010.

[34] M. van Kreveld and B. Speckmann. On rectangular cartograms. *Computational Geometry*, 37(3):175–187, 2007.

[35] M. Krstajić, E. Bertini, and D. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2432–2439, 2011.

[36] M. Krstajić, M. Najm-Araghi, F. Mansmann, and D.A. Keim. Story Tracker: Incremental visual text analytics of news story development. *Information Visualization*, 12(3-4):308–323, 2013.

[37] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[38] J.B. Kruskal and M. Wish. *Multidimensional Scaling*, volume 11. Sage, 1978.

[39] H.A. Le Thi. An efficient algorithm for globally minimizing a quadratic function under convex quadratic constraints. *Mathematical Programming*, 87:401–426, 2000.

[40] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[41] S. Liu, P.-T. Bremer, J.J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):553–562, 2018.

[42] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, 2014.

[43] S. Liu, X. Wang, C. Collins, W. Dou, F. Ouyang, M. El-Assady, L. Jiang, and D. Keim. Bridging text visualization and mining: A task-driven survey. Forthcoming in *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[44] Y. Lu, M. Steptoe, S. Burke, H. Wang, J.-Y. Tsai, H. Davulcu, D. Montgomery, S.R. Corman, and R. Maciejewski. Exploring evolving media discourse through event cueing. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):220–229, 2016.

[45] D. Luo, J. Yang, M. Krstajić, W. Ribarsky, and D. Keim. EventRiver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics*, 18(1):93–105, 2012.

[46] L. vander Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[47] L. vander Maaten and G. Hinton. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55, 2012.

[48] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

[49] A. Marcus, M.S. Bernstein, O. Badar, D.R. Karger, S. Madden, and R.C. Miller. TwitInfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236. ACM, 2011.

[50] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99, 2014.

[51] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[52] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[53] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[54] J. Pennington, R. Socher, and C.D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[55] F. Provost and T. Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc., 2013.

[56] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

[57] S. Rose, S. Butner, W. Cowley, M. Gregory, and J. Walker. Describing story evolution from dynamic information streams. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009.*, pages 99–106, 2009.

[58] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[59] J. Thomas and P.C. Wong. Visual Analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[60] W.S. Torgerson. *Theory and Methods of Scaling.* Wiley, 1958.

[61] F.B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

[62] F.B. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, 2009.

[63] F. Villarroel Ordenes, B. Theodoulidis, J. Burton, T. Gruber, and M. Zaki. Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research*, 17(3):278–295, 2014.

[64] Y. Wang, X. Chu, C. Bao, L. Zhu, O. Deussen, B. Chen, and M. Sedlmair. EdWordle: Consistency-preserving Word Cloud Editing. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):647–656, 2018.

[65] A. Witkin. Physically based modeling: principles and practice constrained dynamics. *Computer Graphics*, 1997.

[66] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. *Computer Graphics Forum*, 30(3):741–750, 2011.

Prof. Emilio Carrizosa

IMUS - Facultad de Matemáticas

Universidad de Sevilla

Sevilla

Spain

Dr. Vanesa Guerrero

Departamento de Estadística

Universidad Carlos III de Madrid

Getafe

Spain

Dr. Daniel Hardt

Department of Digitalization

Copenhagen Business School

Frederiksberg

Denmark

Prof. Dolores Romero Morales

Department of Economics

Copenhagen Business School

Frederiksberg

Denmark

Table 1: Danish words and their translation into English for the `Internet` case study

Danish word	Translation
adressen	address
Apple	Apple
BBS	BBS
bibliotekarer	librarians
browser	browser
cafeen	cafe
CD	CD
computernet	computer network
cyberspace	cyberspace
Dankort	Dankort
danskere	Danes
Diatel	Diatel
diskette	floppy
hackere	hackers
harddisk	hard drive
homepage	homepage
IBM	IBM
Internet_Explorer	Internet Explorer
Jytte_Hilden	Jytte Hilden
KIDLINK	KIDLINK
KØBENHAVN	COPENHAGEN
Mac	Mac
MB	MB
Microsoft	Microsoft
modem	modem
MUD	MUD
Netscape	Netscape
PBS	PBS
PC	PC
Politiken_On_Line	Politiken On Line
Pol_On_Line	Pol On Line
processor	processor
RB	RB
Tele_Danmark	Tele Danmark
Telia	Telia
UNI-C	UNI-C
Windows	Windows
Wired	Wired
World_Wide_Web	World Wide Web
Århus	Århus

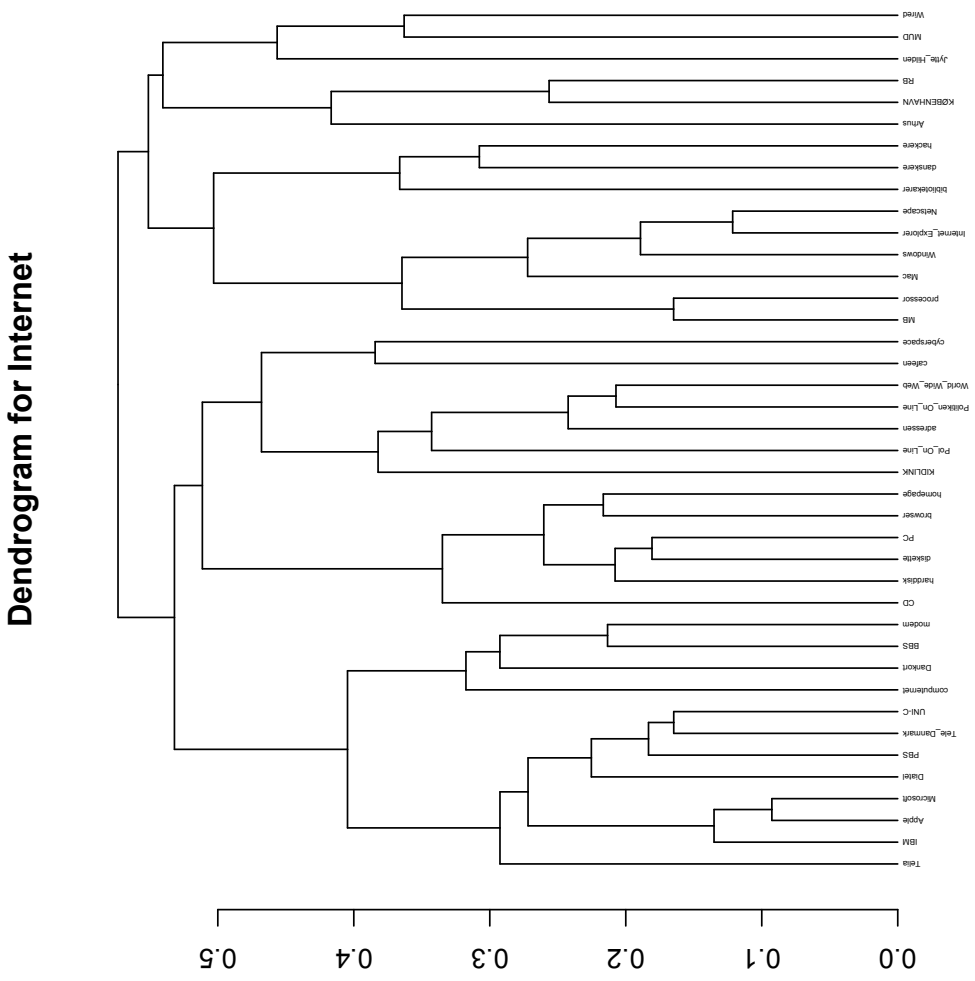

Figure 1: Hierarchical clustering for dissimilarities in the `Internet` case study

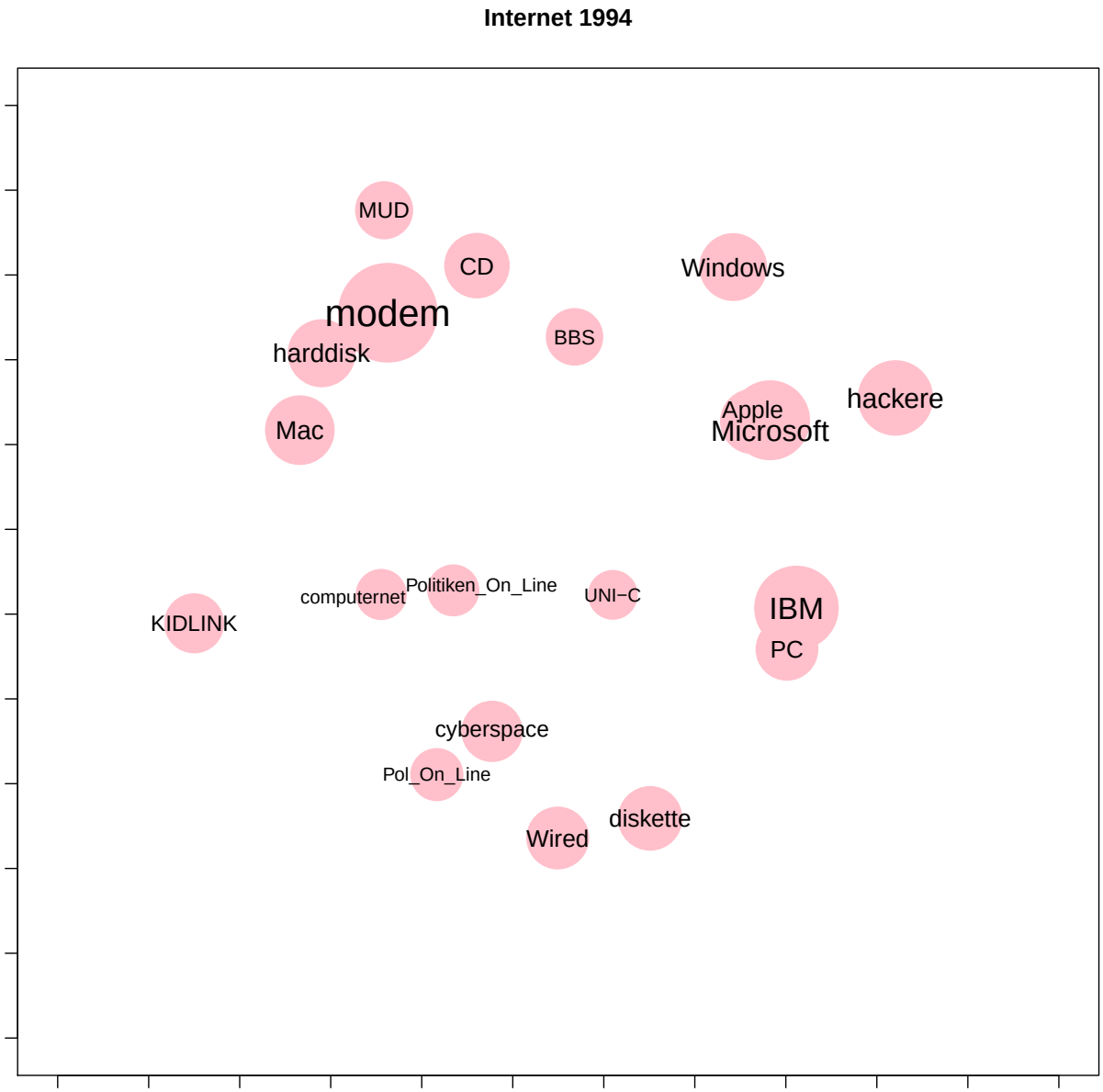

Figure 2: Visualization map for `Internet` in 1994

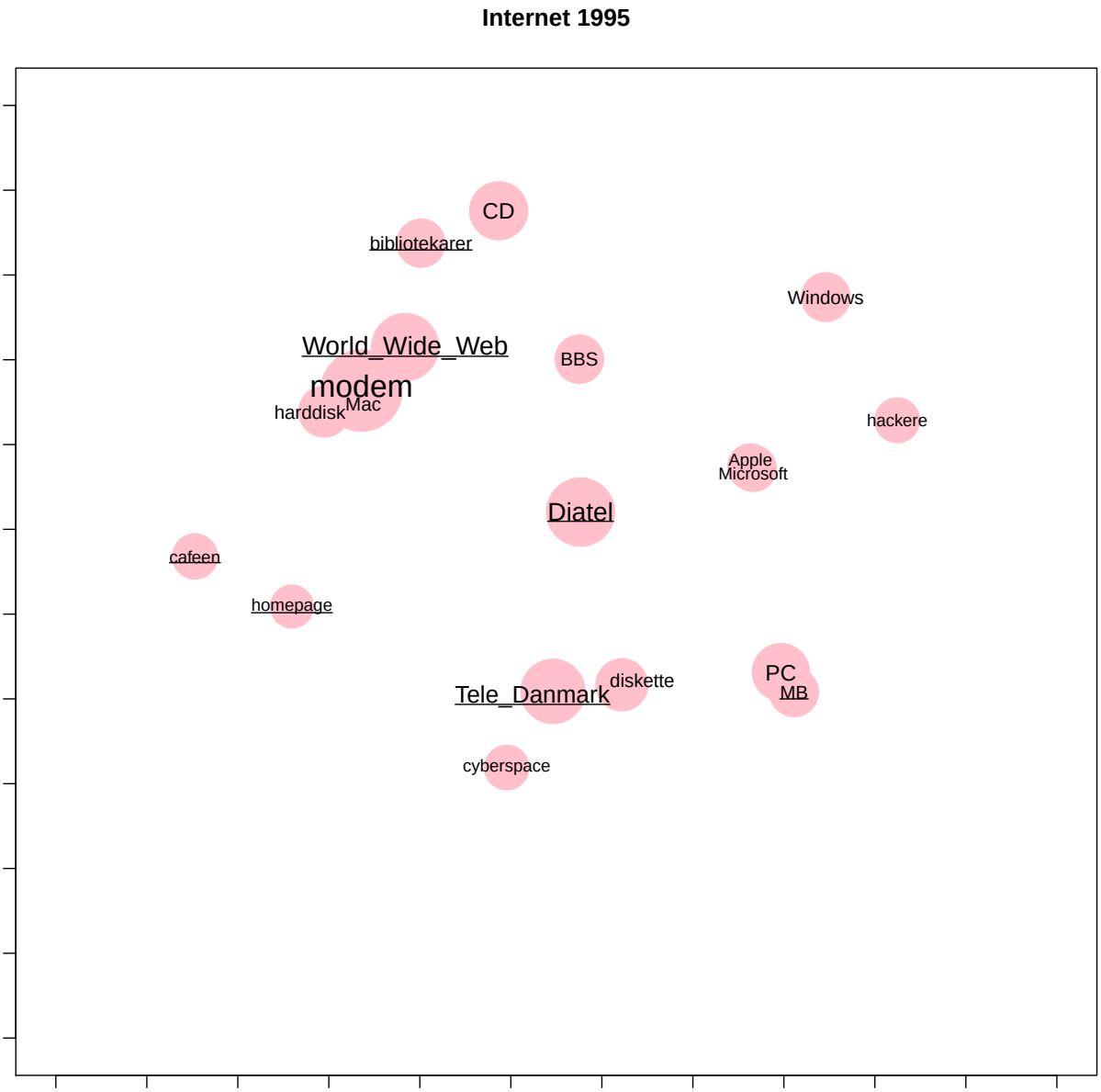

Figure 3: Visualization map for `Internet` in 1995

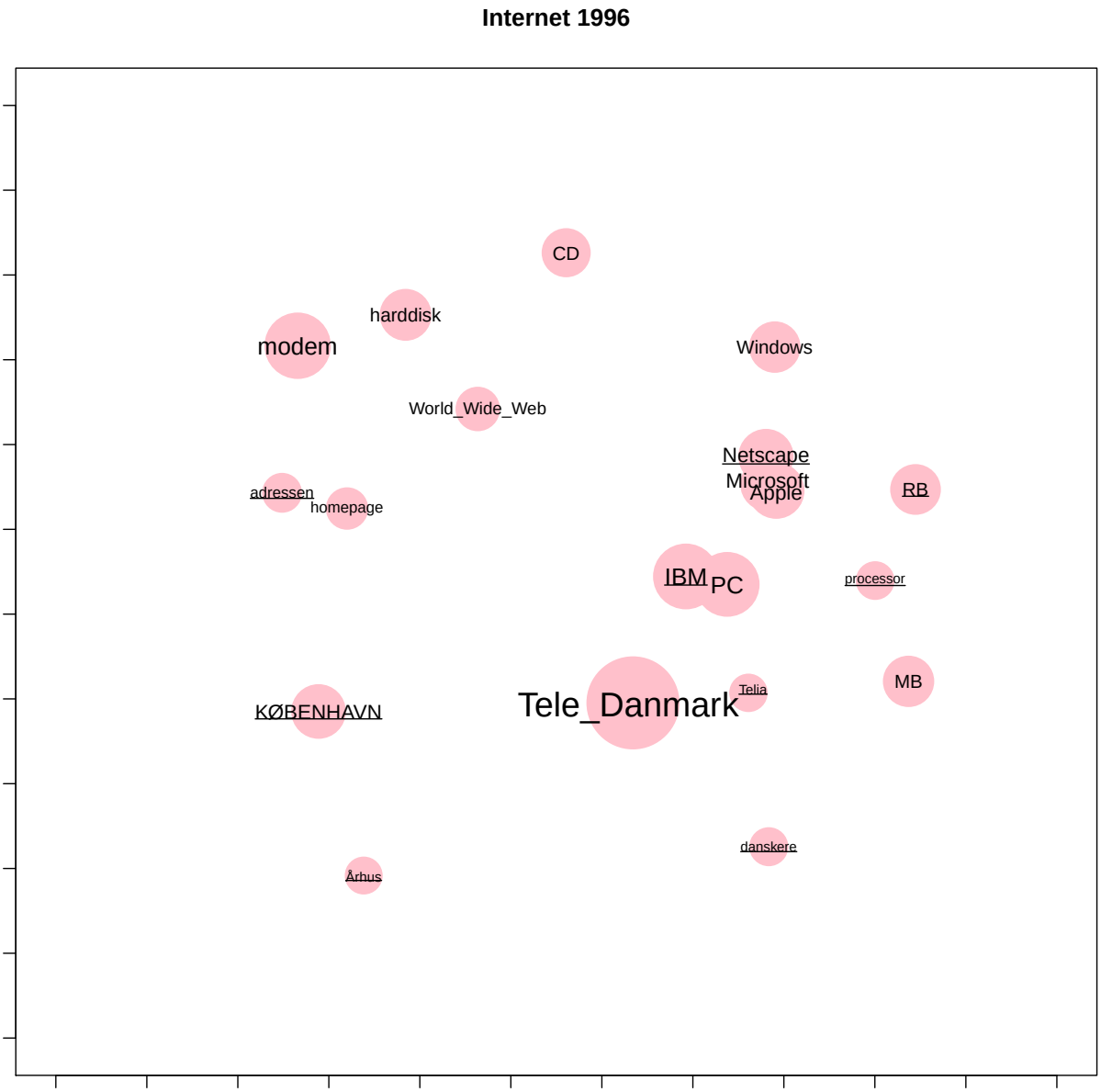

Figure 4: Visualization map for `Internet` in 1996

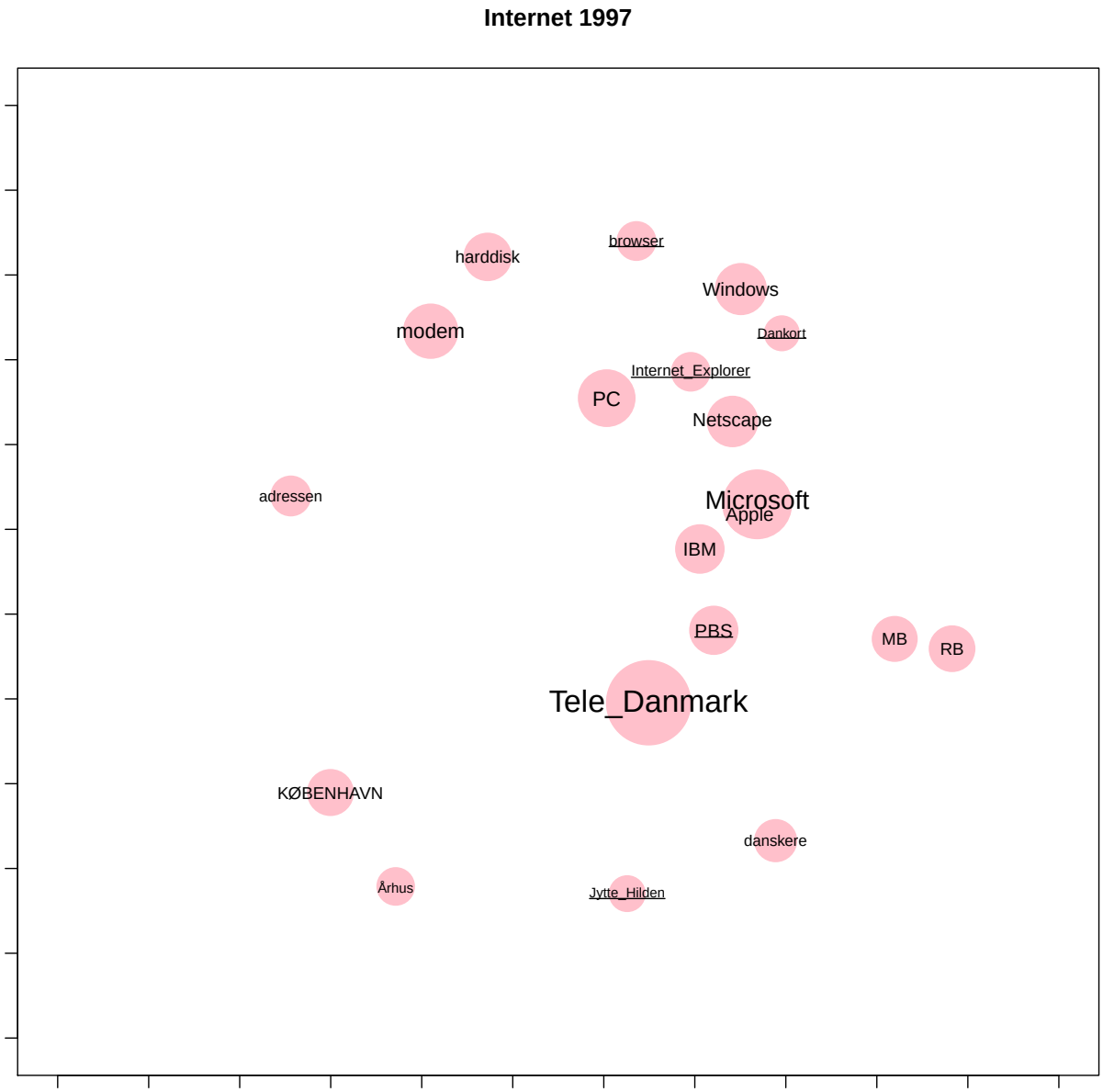

Figure 5: Visualization map for `Internet` in 1997

Table 2: Danish words and their translation into English for the `Terror Attack` case study

Danish word	Translation	Danish word	Translation	Danish word	Translation
AC	AC	HK	HK	republikanerne	Republicans
aftale	appointment	imam	imam	retsordfører	spokesman
aften	evening	Irakkrigen	Iraq War	Reuters	Reuters
Afghanistan	Afghanistan	Islamisk_Stat	Islamic State	Riyadh	Riyadh
al-Awlaki	al-Awlaki	israelske	Israeli	Romney	Romney
al-Qaeda	al-Qaeda	Jens_Stoltenberg	Jens Stoltenberg	SAS	SAS
al-Shabaab	Al-Shabaab	jihadister	jihadists	Saddam	Saddam
al-Zarqawi	al-Zarqawi	John_McCain	John McCain	Said_Mansour	Said Mansour
Alm	Alm	Kabul	Kabul	saudiske	Saudi
Al_Qaida	Al Qaida	Katrina	Katrina	selvmordsbomber	suicide bomber
Amir	Amir	kenyanske	Kenyan	SF	SF
AP	AP	Kerry	Kerry	somaliske	Somali
Bagdad	Baghdad	Krystalgade	Krystalgade	Sotji	Sochi
Bali	Bali	Kurt_Westergaard	Kurt Westergaard	stammeområderne	tribal areas
Beredskabsstyrelsen	Emergency Management	Københavns_Lufthavn	Copenhagen Airport	synagogen	synagogue
Beslan	Beslan	Lars	Lars	Taleban	Taliban
BH	BH	LH	LH	tegninger	drawings
bin	bin	lufthavn	airport	terrorangreb	terrorist attacks
bombemænd	bomb men	Manhattan	Manhattan	terrorbomberne	bombings
Boston	Boston	masseødelæggelsesvåben	WMD	terrormistænkte	terror suspects
Breivik	Breivik	Maximus	Maximus	Twitter	Twitter
Bush	Bush	Mette_Frederiksen	Mette Frederiksen	uighurer	Uighurs
Charlie_Hebdo	Charlie Hebdo	miltbrand	anthrax	Utøya	Utøya
Chicago	Chicago	Mombasa	Mombasa	Villy_Søvndal	Villy Søvndal
CIA	CIA	Moore	Moore	Volgograd	Volgograd
Clarke	Clarke	Morten_Storm	Morten Storm	World_Trade_Center	World Trade Center
danskere	Danes	moske	mosque	Yemen	Yemen
Dansk_Folkeparti	Danish People's Party	Muhammed	Muhammed		
Det_Hvide_Hus	The White House	Mumbai	Mumbai		
DF	DF	Musharraf	Musharraf		
droneangreb	drone attacks	muslimer	Muslims		
Edward_Snowden	Edward Snowden	New_Orleans	New Orleans		
efterfølgende	subsequent	norske	Norwegian		
efterretningstjenester	intelligence services	NSA	NSA		
eftektivt	effectively	nyhedsbureauet	news agency		
efterretningstjeneste	intelligence service	Nyrup	Nyrup		
el-Hussein	el-Hussein	Obama	Obama		
el-Sheikh	el-Sheikh	offentlige	public		
Facebook	Facebook	Omar_El-Hussein	Omar El-Hussein		
FBI	FBI	Osama	Osama		
flykaprere	hijackers	Oslo	Oslo		
Fogh	Fogh	pakistanske	Pakistani		
forbundspoliti	Federal police	Pentagon	Pentagon		
gasanlægget	gas plant	PET	PET		
Giuliani	Giuliani	PFA	PFA		
Ground_Zero	Ground Zero	profeten	Prophet		
Hamas	Hamas	præsidentkandidat	presidential candidate		
Headley	Headley	PST	PST		
Helle_Thorning-Schmidt	Helle Thorning-Schmidt	radikaliserede	radicalized		
Hizbollah	Hezbollah	Rana	Rana		

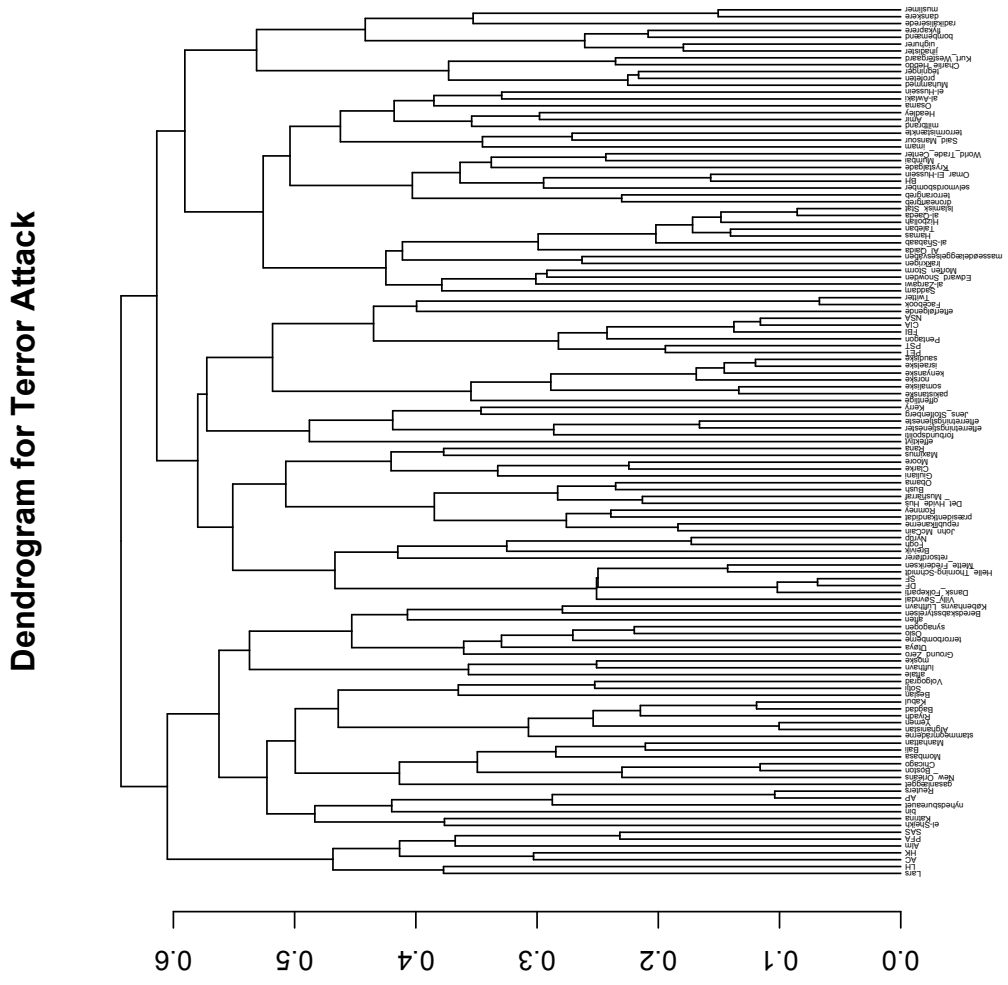

Figure 6: Hierarchical clustering for dissimilarities in the `Terror Attack` case study

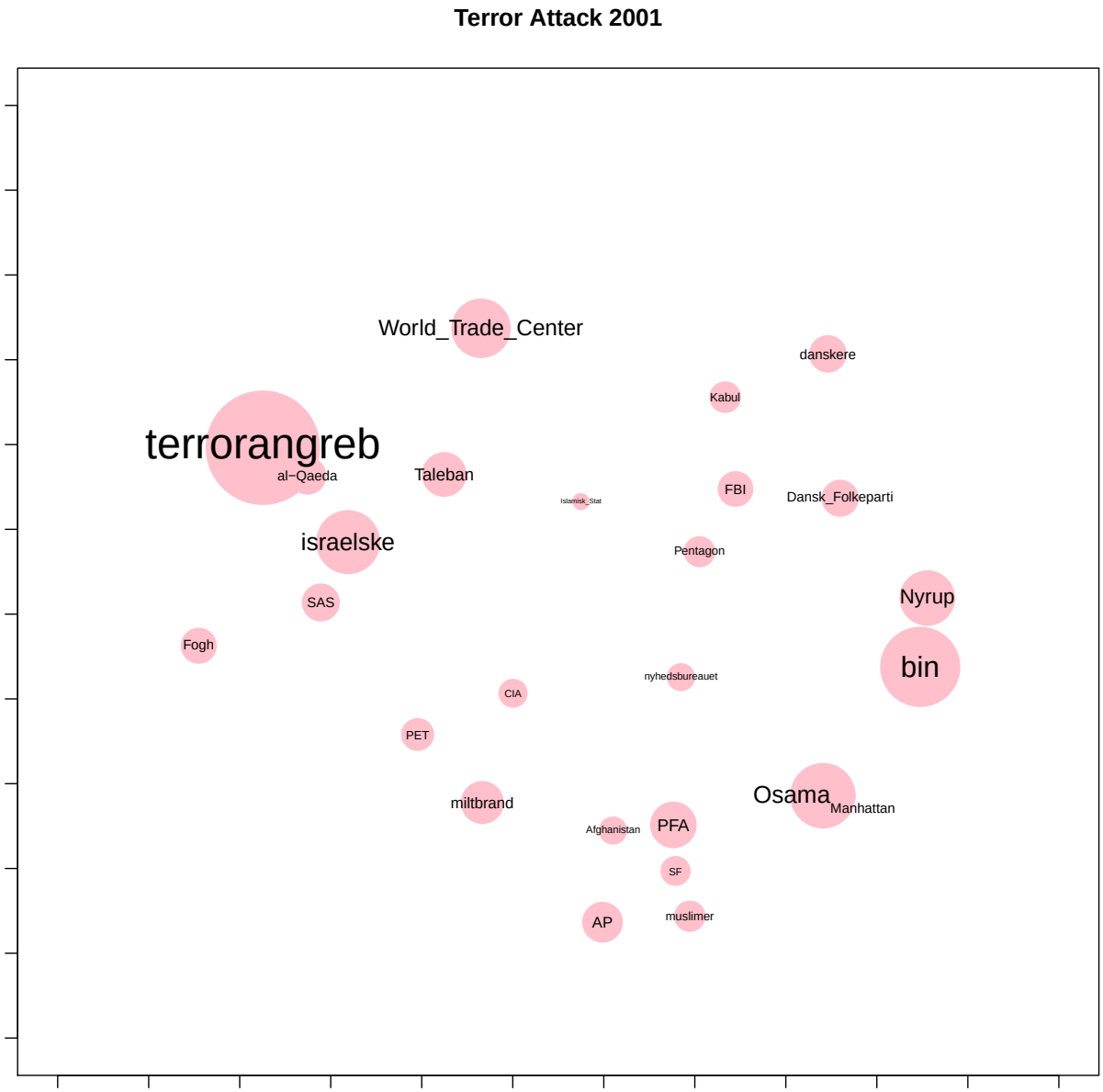

Figure 7: Visualization map for `Terror Attack` in 2001

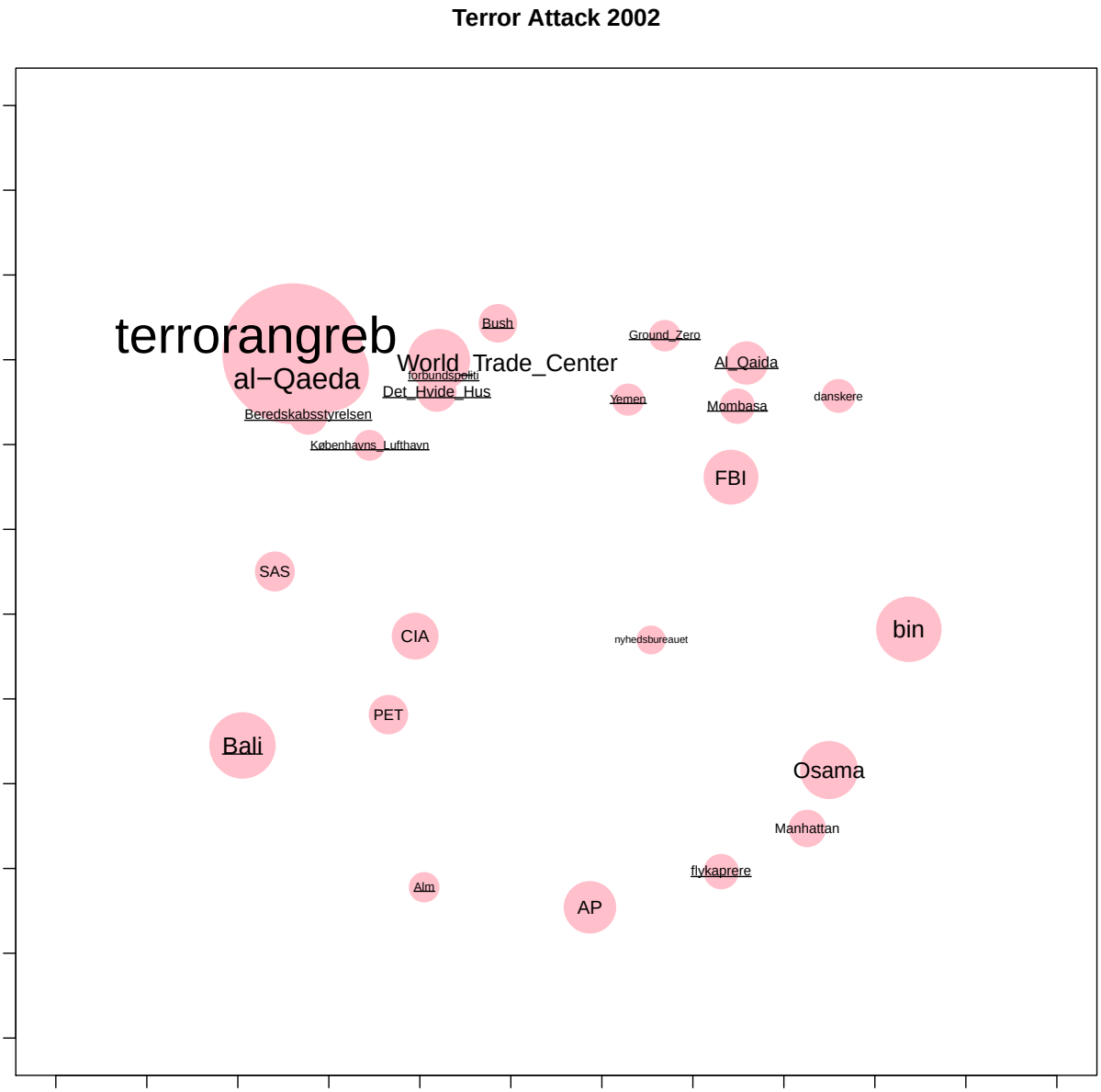

Figure 8: Visualization map for `Terror Attack` in 2002

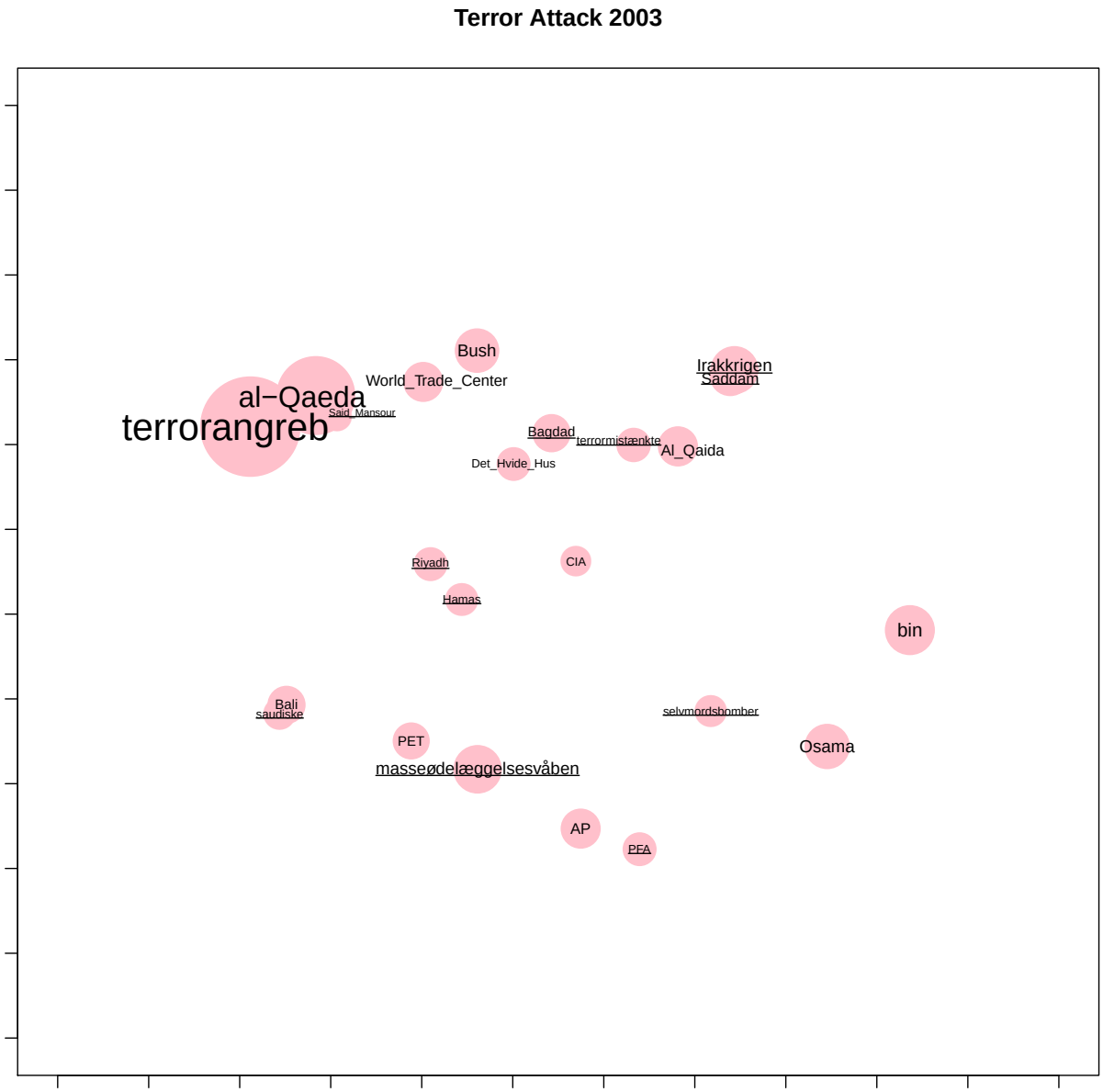

Figure 9: Visualization map for `Terror Attack` in 2003

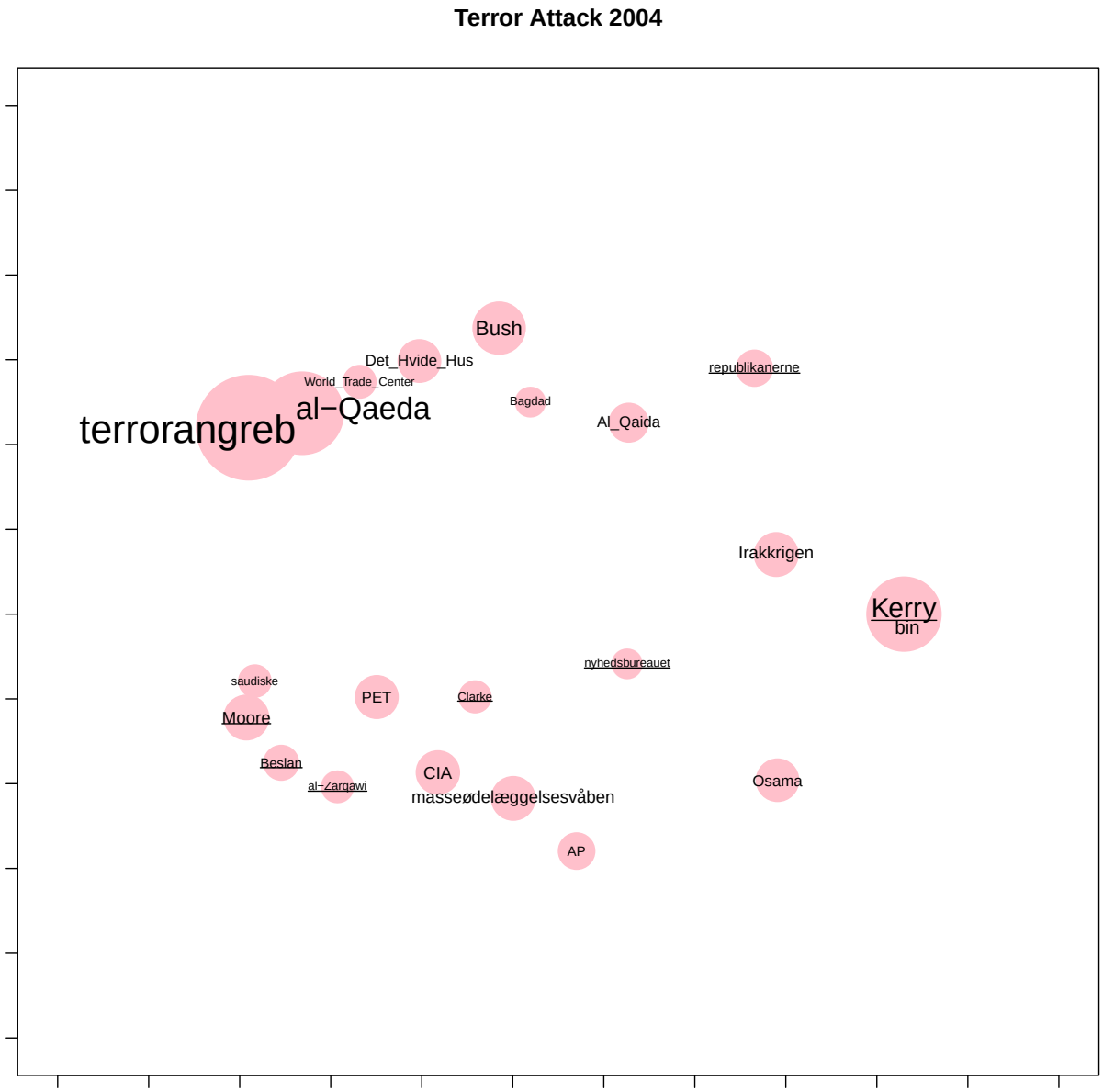

Figure 10: Visualization map for `Terror Attack` in 2004

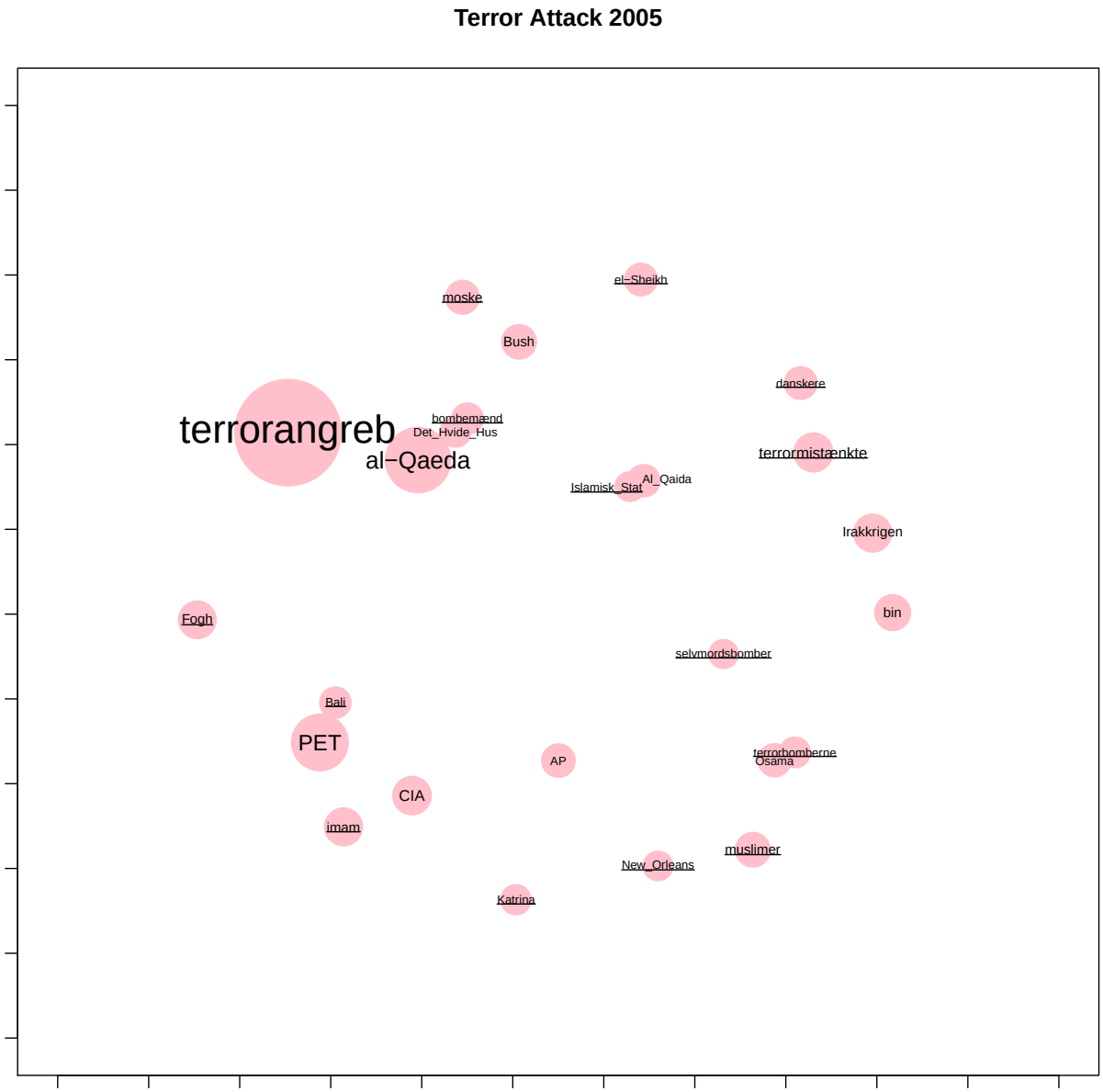

Figure 11: Visualization map for `Terror Attack` in 2005

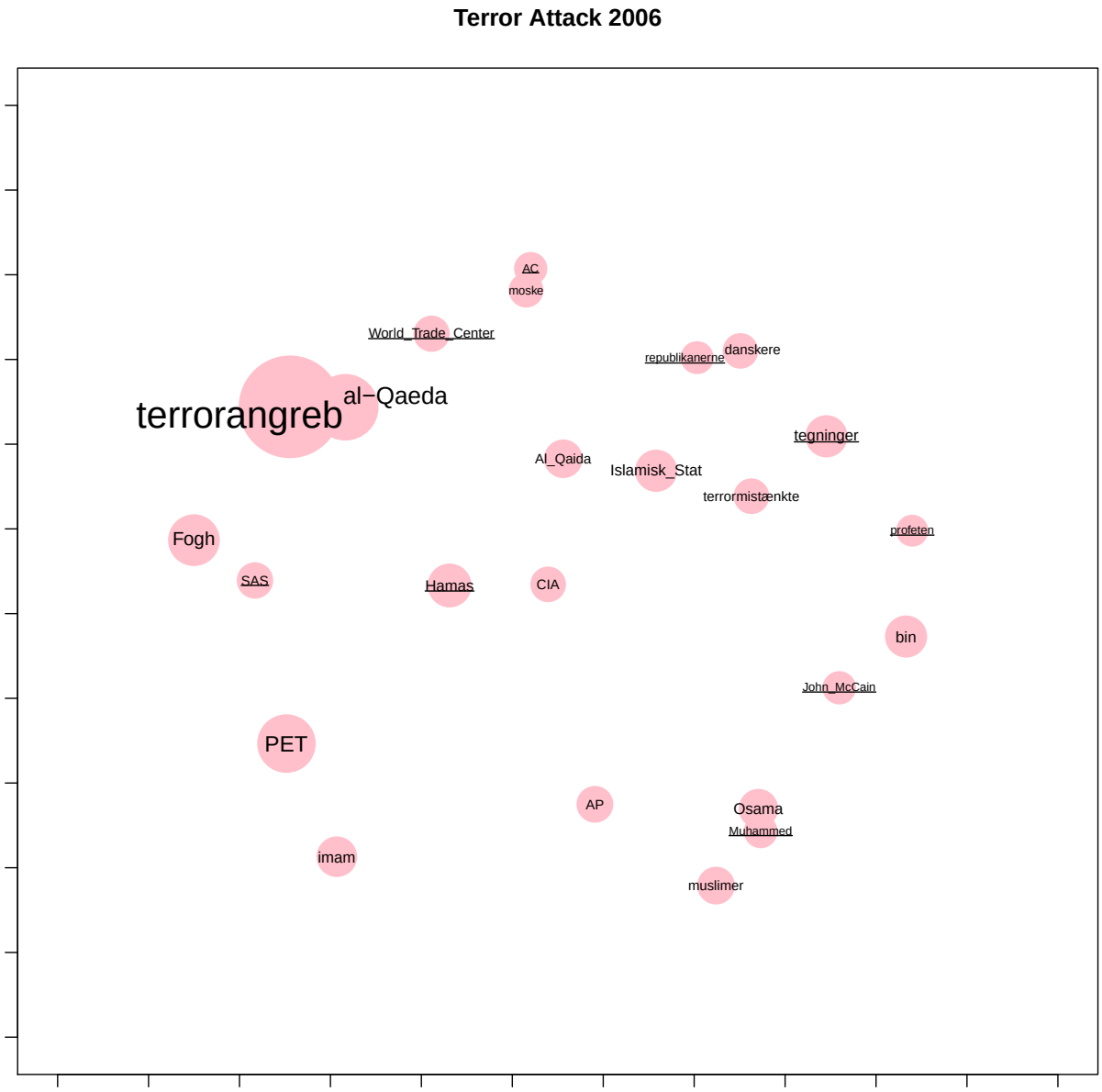

Figure 12: Visualization map for `Terror Attack` in 2006

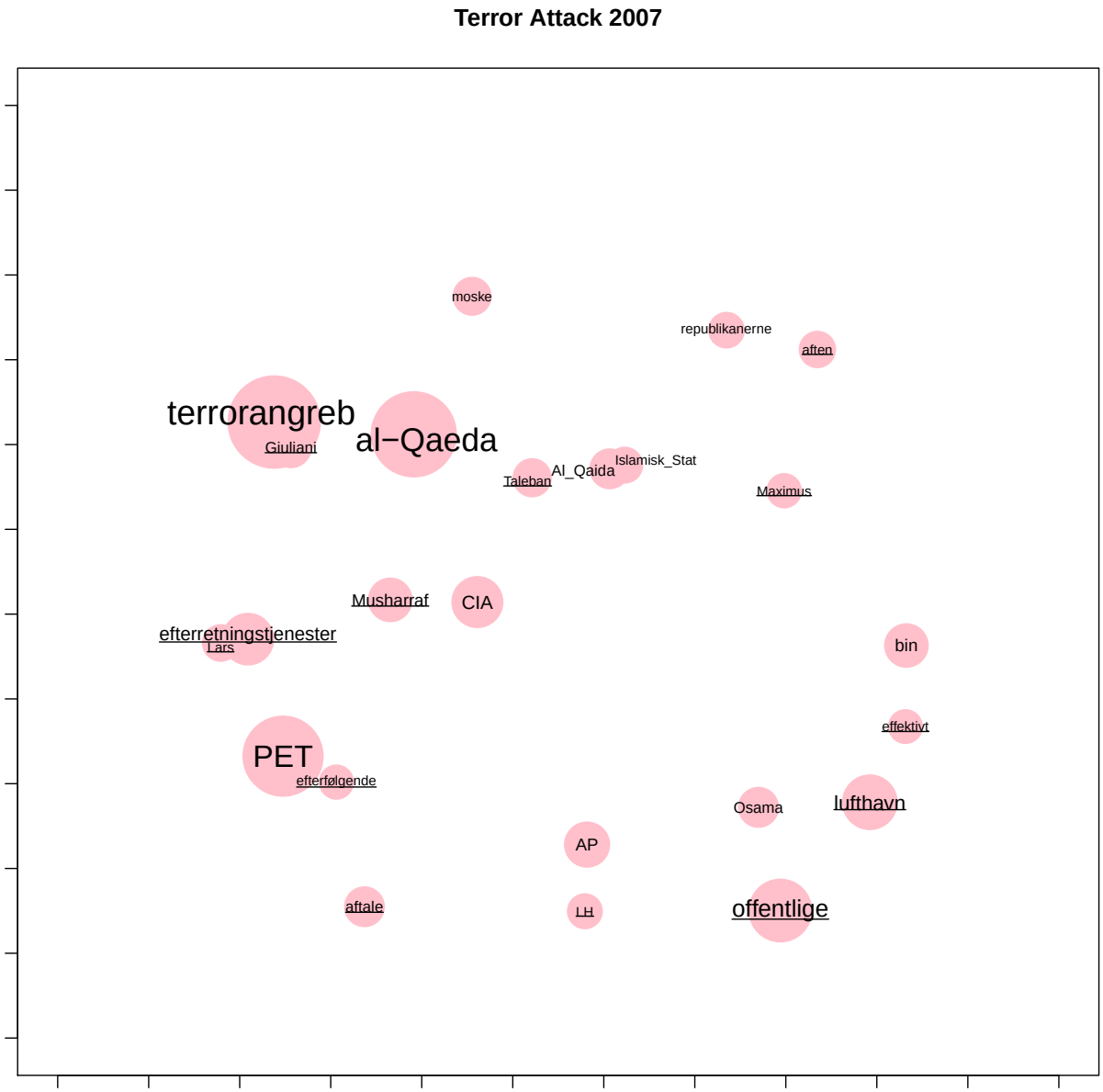

Figure 13: Visualization map for `Terror Attack` in 2007

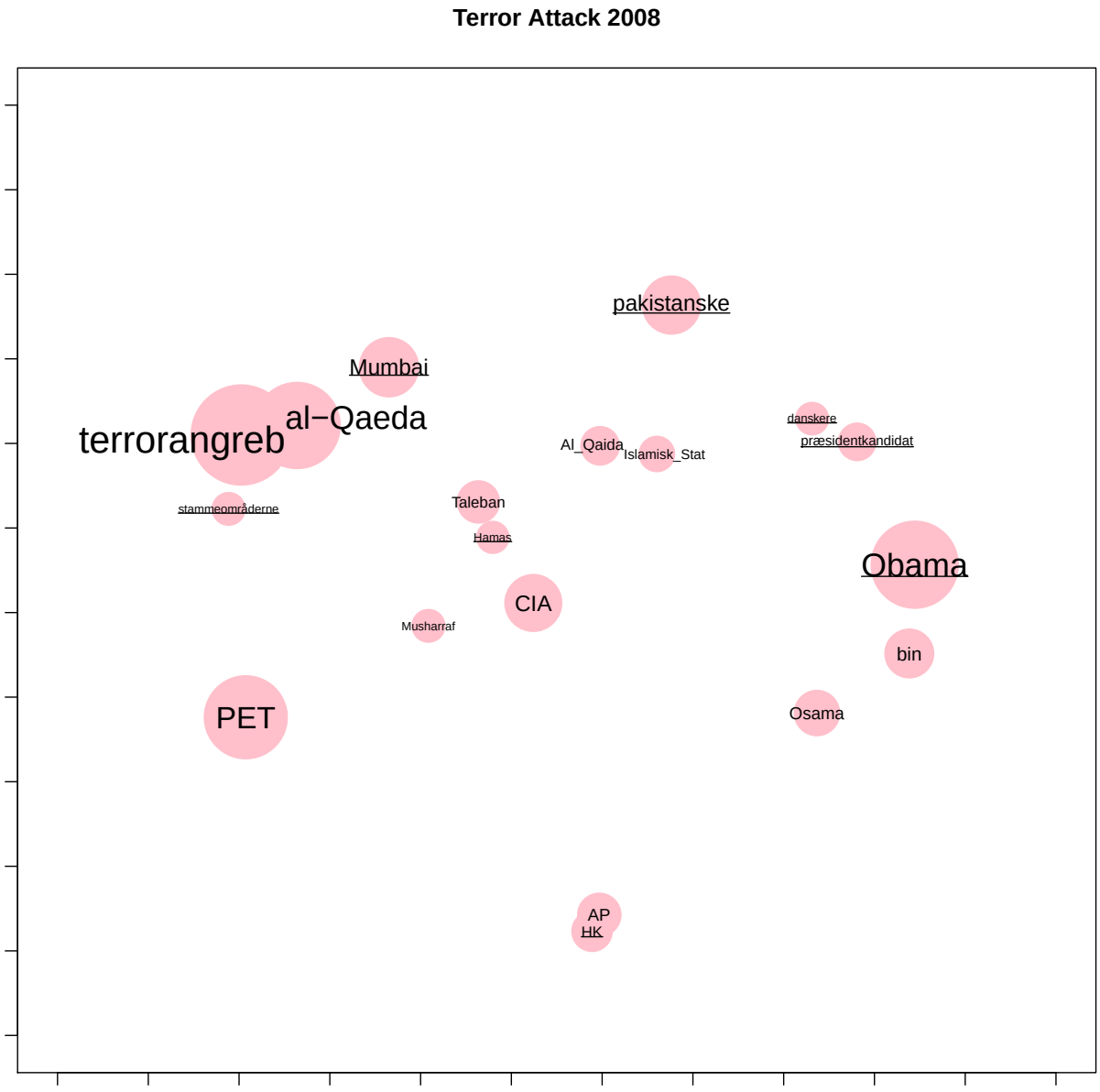

Figure 14: Visualization map for `Terror Attack` in 2008

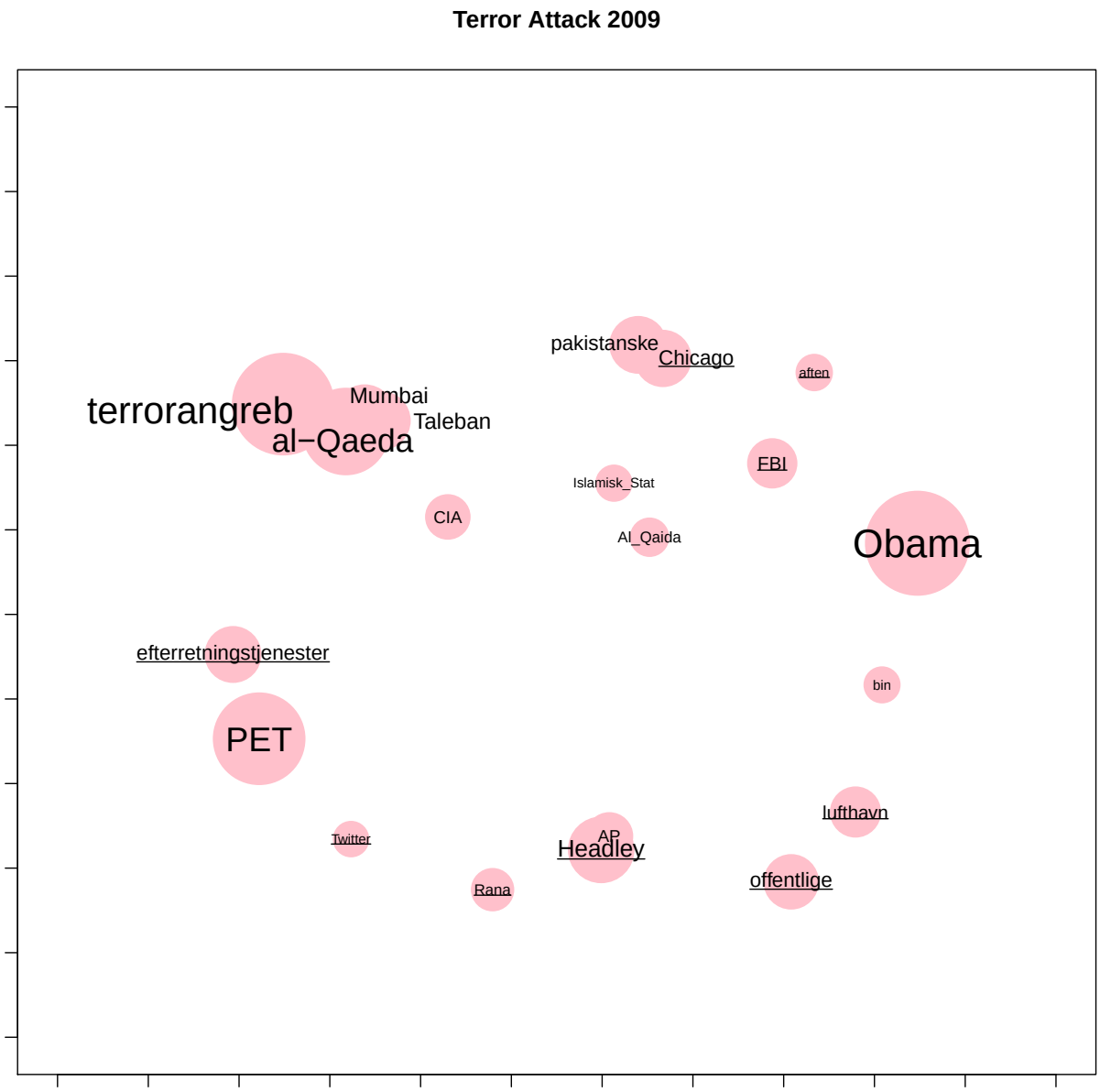

Figure 15: Visualization map for `Terror Attack` in 2009

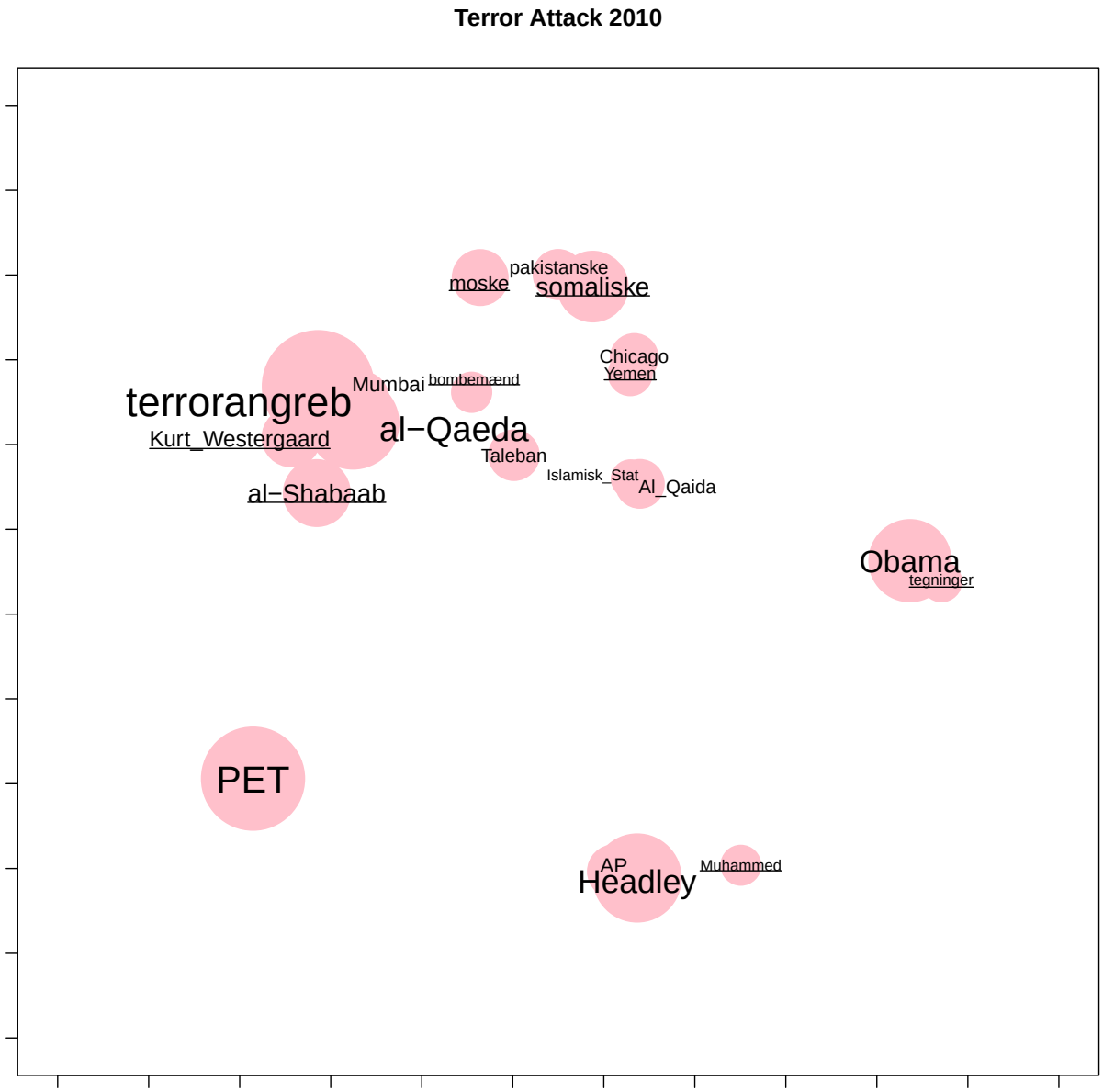

Figure 16: Visualization map for `Terror Attack` in 2010

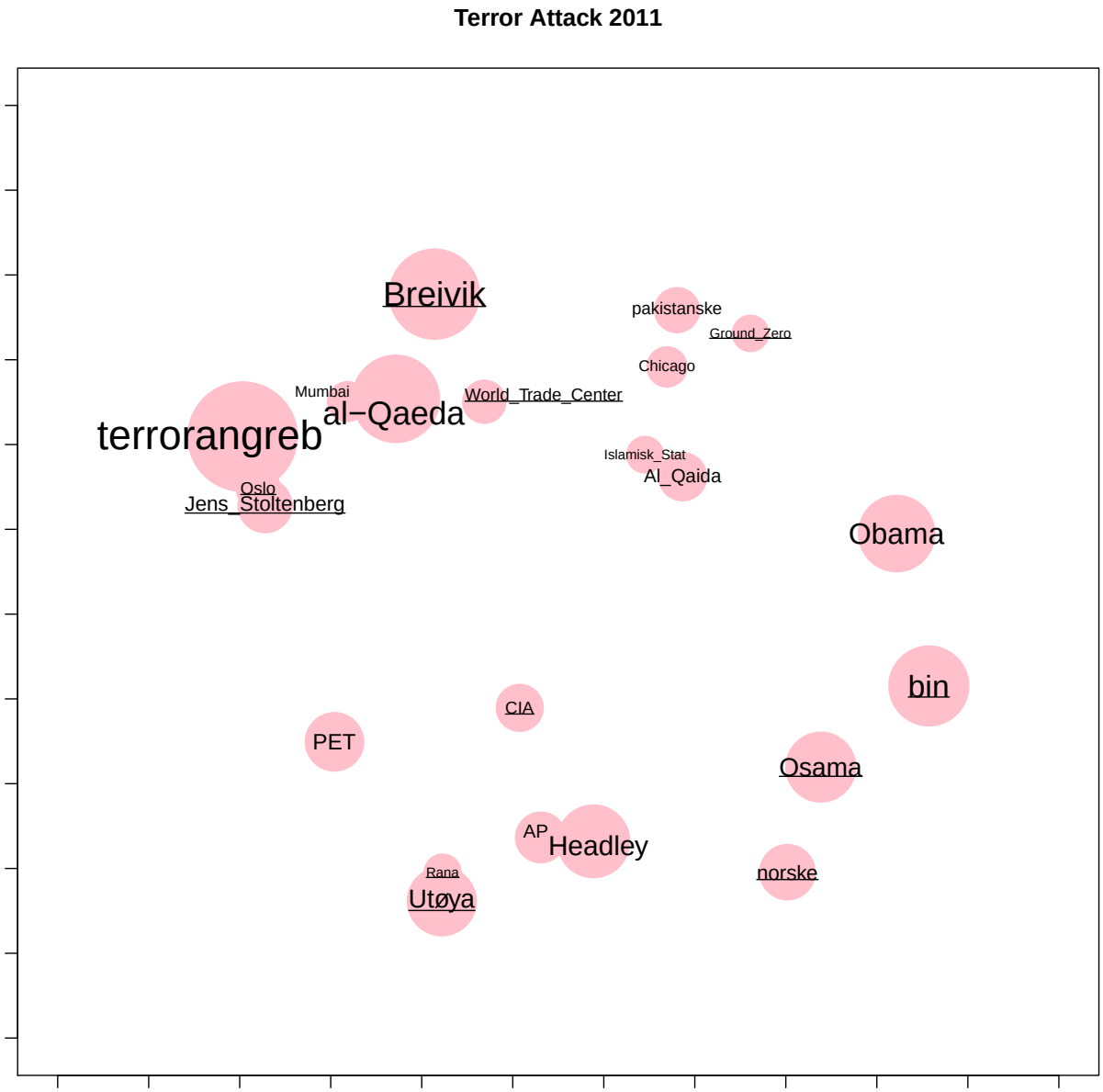

Figure 17: Visualization map for `Terror Attack` in 2011

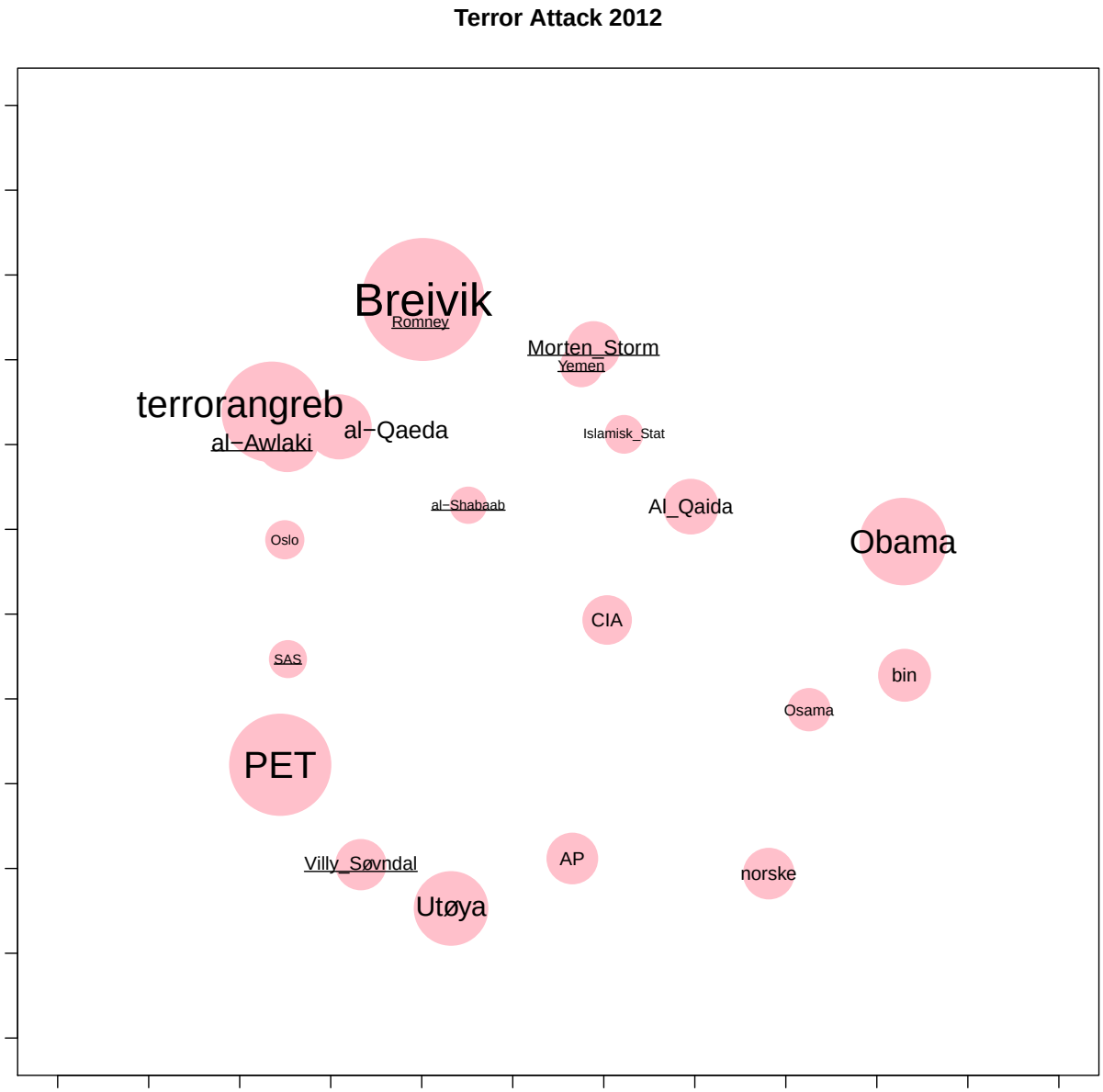

Figure 18: Visualization map for `Terror Attack` in 2012

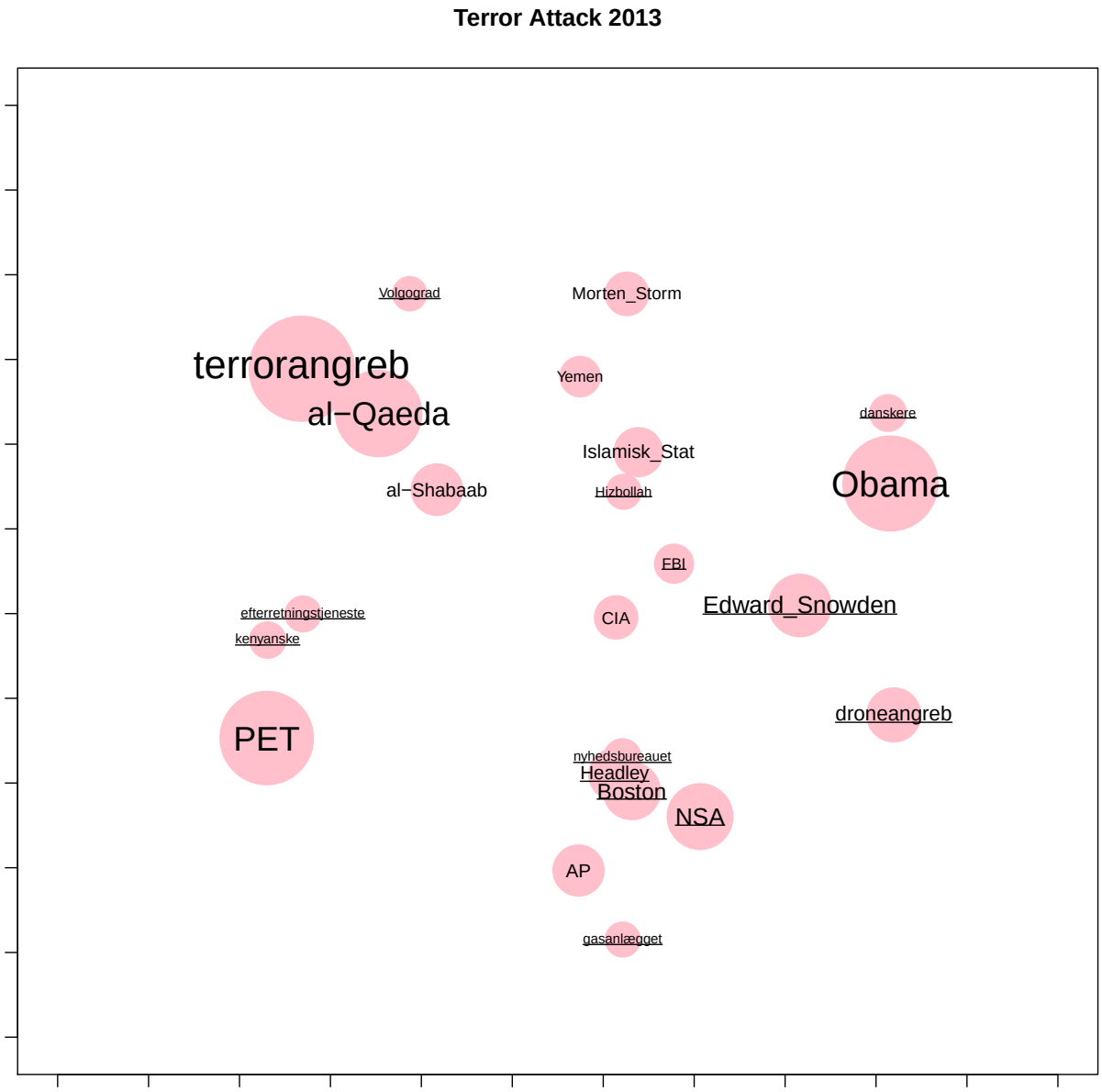

Figure 19: Visualization map for `Terror Attack` in 2013

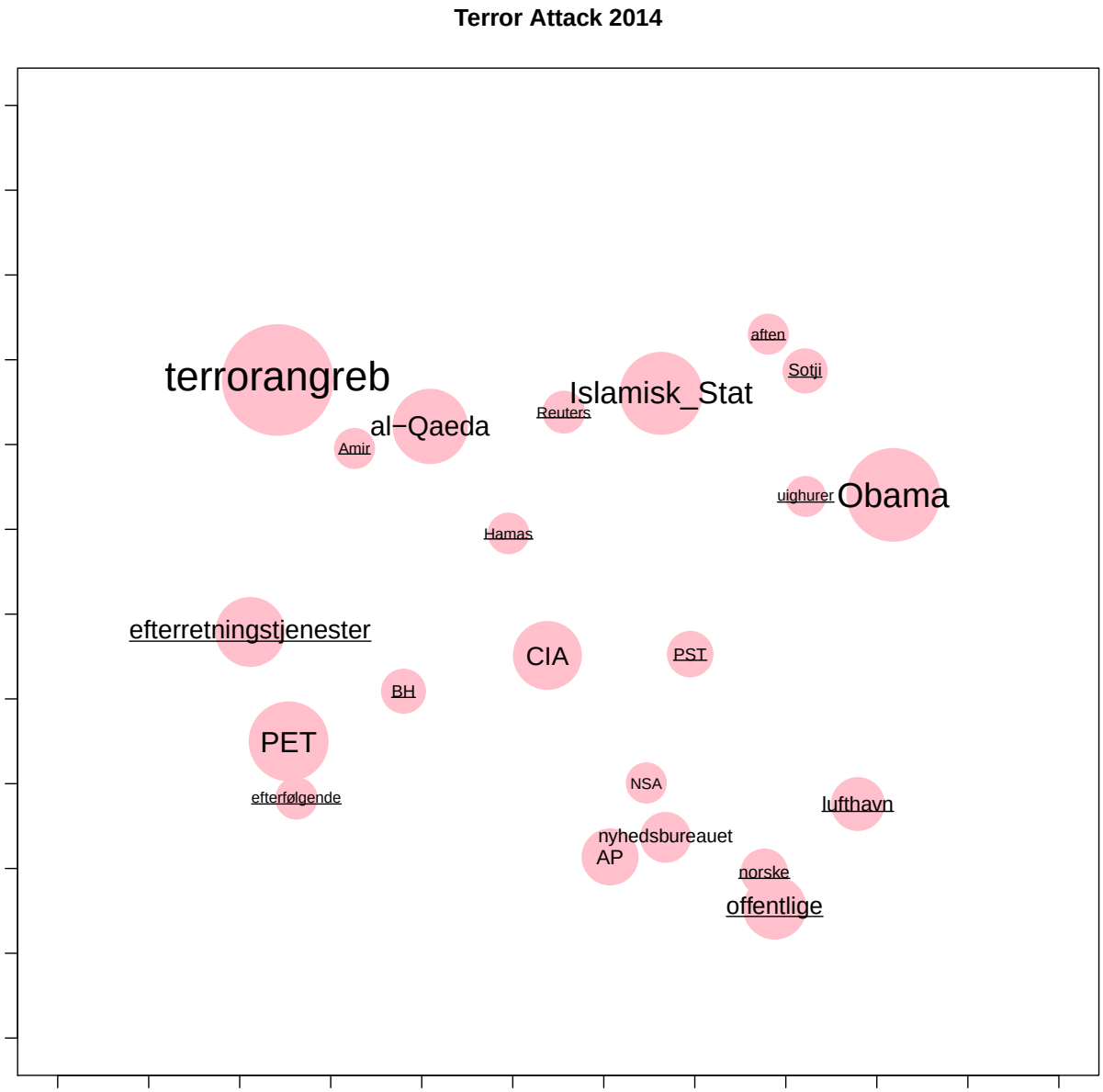

Figure 20: Visualization map for `Terror Attack` in 2014

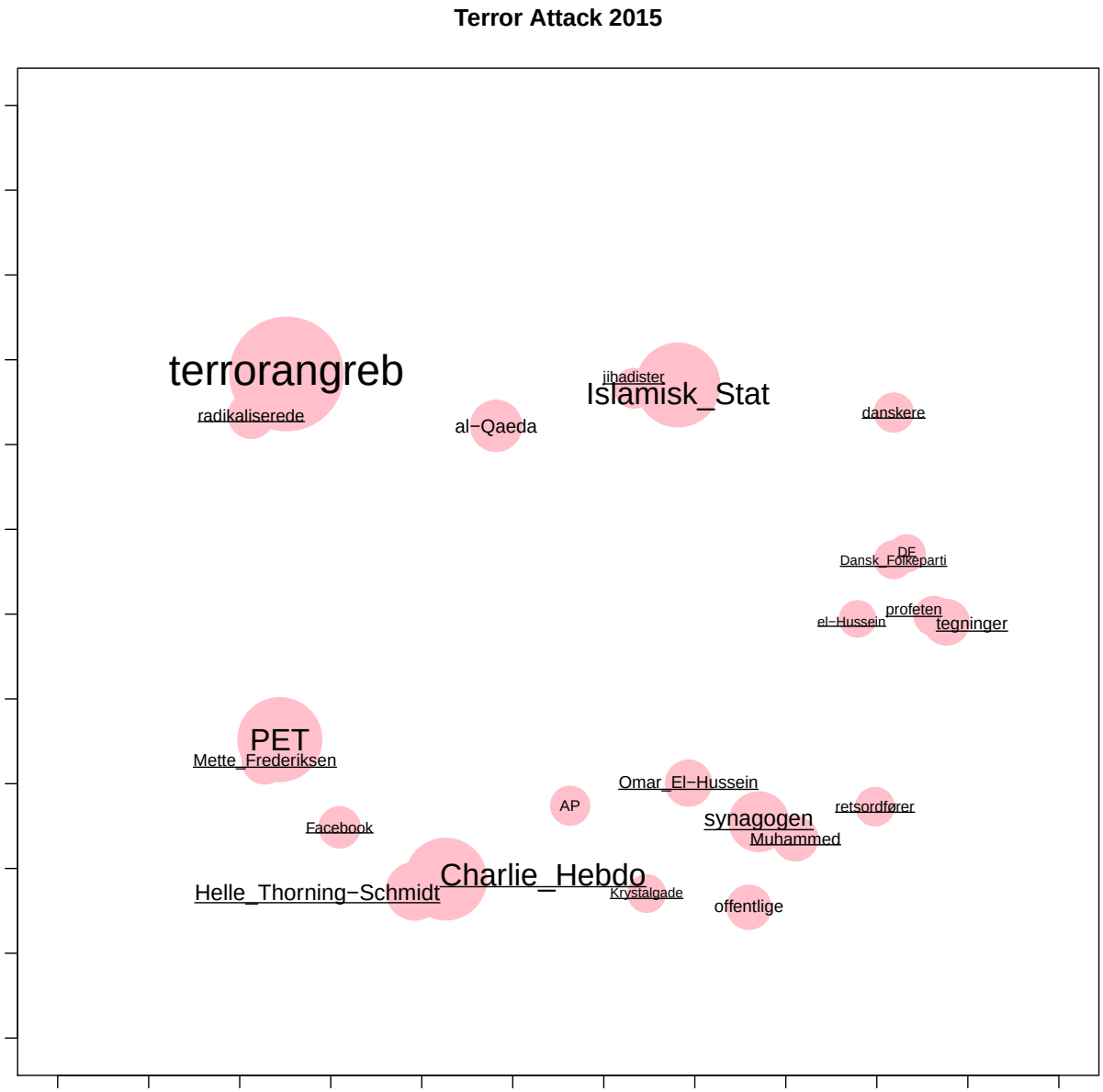

Figure 21: Visualization map for `Terror Attack` in 2015

Table 3: Danish words and their translation into English for the `Immigration` case study

Danish word	Translation	Danish word	Translation	Danish word	Translation	Danish word	Translation
Aarhus	Aarhus	Frank_Jensen	Frank Jensen	Københavns_Universitet	University of Copenhagen	Romney	Romney
Abu_Laban	Abu Laban	Frederik	Frederik	Lars_Hedegaard	Lars Hedegaard	Rosengård	Rosengård
Ali	Ali	fremmede	foreign	Lars_Løkke_Rasmussen	Lars Løkke Rasmussen	sammenhængskraft	cohesion
andengenerationsindvandrere	second generation immigrant	Fremskridtspartiet	Progress Party	Leo	Leo	Sarkozy	Sarkozy
Anne_Knudsen	Anne Knudsen	Fryd	Joy	Leyla	Leyla	SF	SF
Arizona	Arizona	Fædregruppen	Fathers Group	LO	LO	sgu	damn
arj	arj	Gellerup	Gellerup	Louise_Gade	Louise Gade	sjællænder	Sealander
bandekrigen	gang war	ghettoer	ghettos	Lubna_Elahi	Lubna Elahi	skuespiller	actor
Basim	Basim	Glistrup	Glistrup	Madsen	Madsen	skyderier	shootings
beduinerne	Bedouins	Gogh	Gogh	Majid	Majid	Sleiman	Sleiman
Berlingske_Tidende	Berlingske Tidende	Grønland	Greenland	Malmø	Malmo	Socialdemokratiet	Social Democratic Party
Bertel_Haarder	Bertel Haarder	Gyldent_Daggry	Gyldent Daggry	Mamadou	Mamadou	socialrådgiver	social worker
Birte_Weiss	Birte Weiss	halalhippier	halal hippies	Manu_Sareen	Manu Sareen	somaliere	Somalis
Bispehaven	Bispehaven	Halle	Halle	Marianne_Jelved	Marianne Jelved	stationen	station
Bjørn_Nørgaard	Bjørn Nørgaard	Hamid_Rahmati	Hamid Rahmati	Marwan	Marwan	Strauss-Kahn	Strauss-Kahn
blog	blog	Hans_Lassen	Hans Lassen	Mazhar_Hussain	Mazhar Hussain	stuepige	maid
Blågårds_Plads	Blågårds Plads	Hasan	Hasan	McCain	McCain	Suleebo	Suleebo
bogen	the book	Hells_Angels	Hells Angels	menighed	congregation	Sverigedemokraterne	Sweden Democrats
Brian	Brian	Henrik	Henrik	mentor	mentor	Søren_Pind	Søren Pind
Bycyklen	City Bike	herberg	hostelry	MFP	MFP	taxa	taxi
bydelsrådet	town council	hvidløg	garlic	Mir	Mir	tegninger	drawings
bæveren	beaver	Hüseyin_Arac	Hüseyin Arac	Mogens_Camre	Mogens Camre	Theo	Theo
Carlos	Carlos	ideer	ideas	moskeer	mosques	Thorkild_Simonsen	Thorkild Simonsen
Charlie_Hebdo	Charlie Hebdo	ikke-vestlige	non-Western	Muhammed	Muhammed	Thorvaldsen	Thorvaldsen
Chopin	Chopin	imamer	imams	museum	museum	Tingbjerg	Tingbjerg
Cirkeline	Cirkeline	IND-sam	IND-SAM	muslimer	Muslims	torvet	square
cyklerne	bikes	indvandrerbaggrund	immigrant background	Mustafa	Mustafa	tosprogede	bilingual
Danmarks_Statistik	Statistics Denmark	Inger_Støjberg	Inger Støjberg	mårhunde	raccoon dogs	tvangsægteskaber	forced marriages
danskere	Danes	Instr	Director	Nakskov	Nakskov	tyrkere	Turks
Dansk_Folkeparti	Danish People?s Party	integrationsminister	Minister of integration	Ny_Alliance	Ny Alliance	tørklæde	scarf
danskundervisning	danish lessons	Ishøj	Ishøj	Nyrup	Nyrup	udlændingepolitik	immigration policy
DF	DF	islam	Islam	nævningene	jury	uighurer	Uighurs
Dieudonné	Dieudonné	Islamisk_Stat	Islamic State	Nørrebro	Nørrebro	Ulla_Dahlerup	Ulla Dahlerup
digte	poems	Jamal	Jamal	Obama	Obama	ulve	wolves
dingoen	dingo	Jeppe	Jeppe	Odense	Odense	urfuglen	grouse
drabsforsøg	attempted murder	Josef	Josef	OL	Olympic Games	vampyr	vampire
drenge	boys	jyder	Jutlanders	opgangen	entryway	Vangsgaard	Vangsgaard
DSK	DSK	jyske	Native Jutland	ORG	ORG	velfærdsturisme	welfare tourism
dørmænd	doormen	Jønke	Jønke	pakistanere	Pakistanis	Venstre	Venstre
ECRI	ECRI	Karen_Jespersen	Karen Jespersen	PET	PET	Vesterbro	Vesterbro
Egtvedpigen	Egtvedpigen	Kassem	Kassem	PhD	PhD	vildsvin	boar
Ekstra_Bladet	Ekstra Bladet	Kenneth	Kenneth	Pia Kjærsgaard	Pia Kjærsgaard	Villy_Søvndal	Villy Søvndal
elg	moose	Khader	Khader	Pim_Fortuyn	Pim Fortuyn	Vollsmose	Vollsmose
Esbjerg	Esbjerg	Khaled_Ramadan	Khaled Ramadan	POEM	POEM	Wallait_Khan	Wallait Khan
Eske	Eske	kontanthjælp	cash assistance	profeten	Prophet	Wilders	Wilders
Facebook	Facebook	Krasnik	Krasnik	Rabin	Rabin	Yahya_Hassan	Yahya Hassan
familiesammenførte	reunited family	Krim	Crimea	RB	RB	Århus	Aarhus
filmen	film	Kristeligt_Folkeparti	Kristeligt Folkeparti	Reich	Reich	Århusianske	Native Aarhus
flygel	Grand piano	Kristian_Thulesen_Dahl	Kristian Thulesen Dahl	Republikanerne	Republicans	Özlem	Özlem
Fogh	Fogh	kronik	feature article	rockere	rockers		
Folketinget	the Danish parliament	Kurt	Kurt	romaer	Roma		

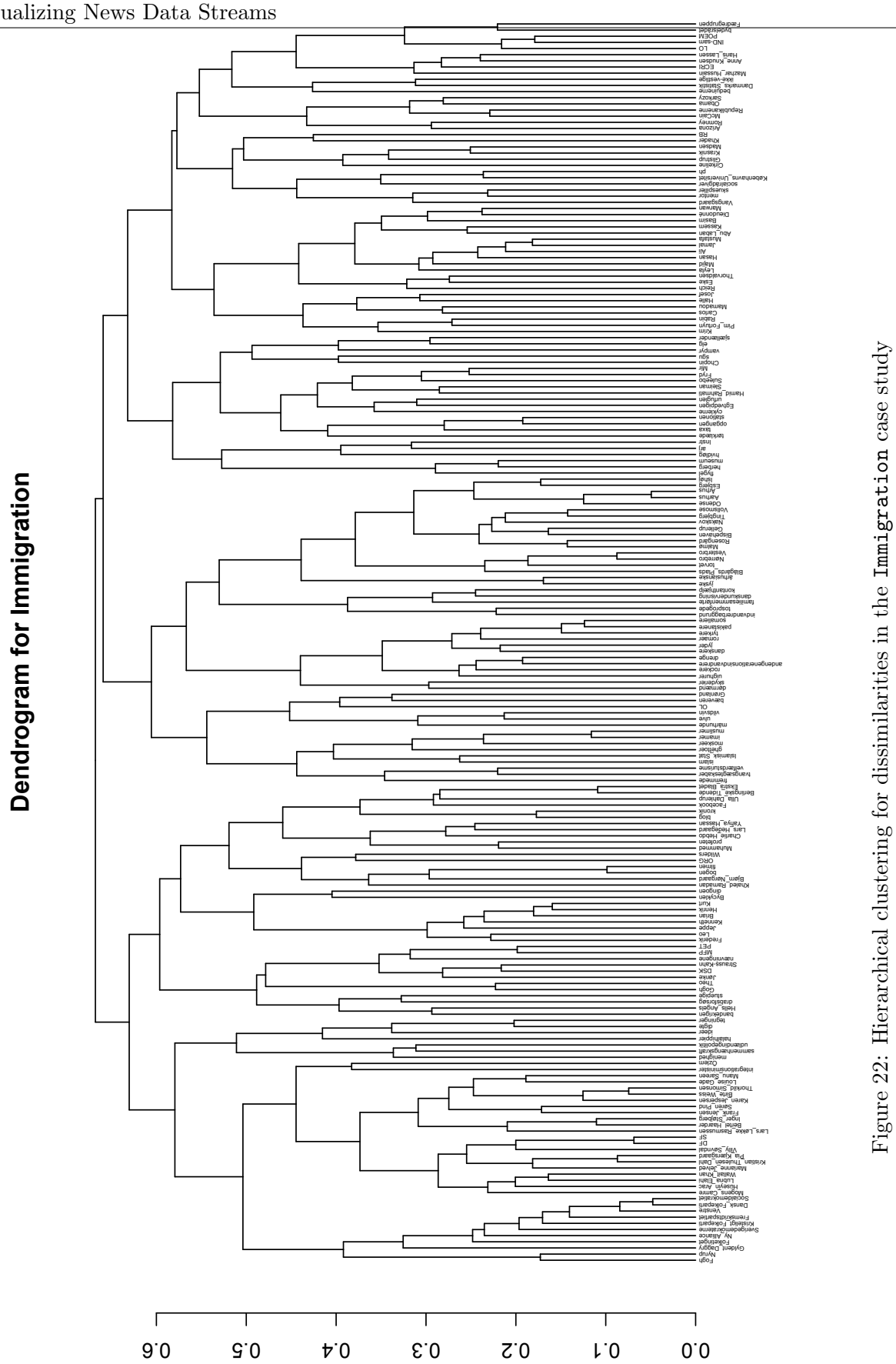

Figure 22: Hierarchical clustering for dissimilarities in the `Immigration` case study

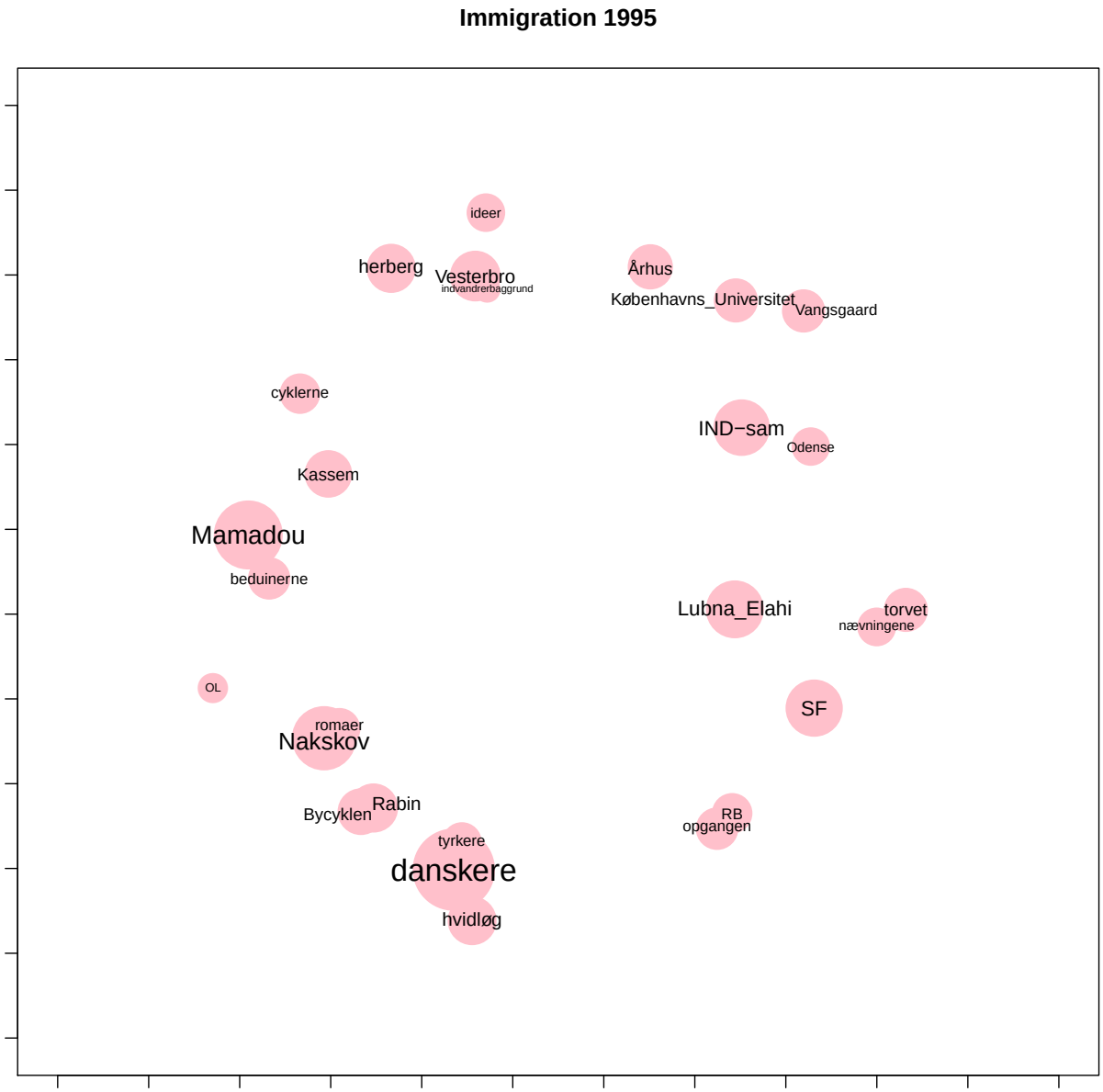

Figure 23: Visualization map for `Immigration` in 1995

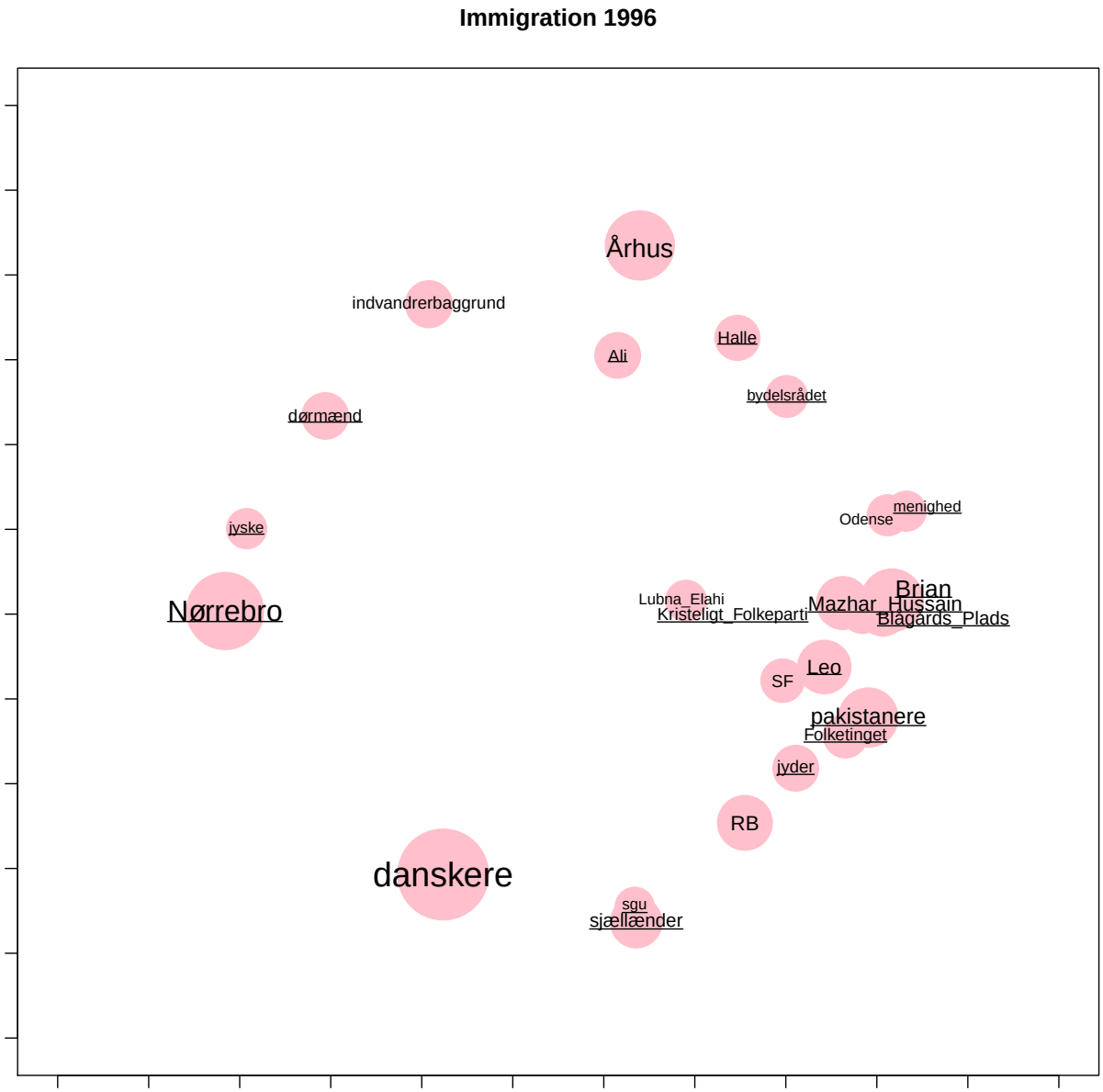

Figure 24: Visualization map for `Immigration` in 1996

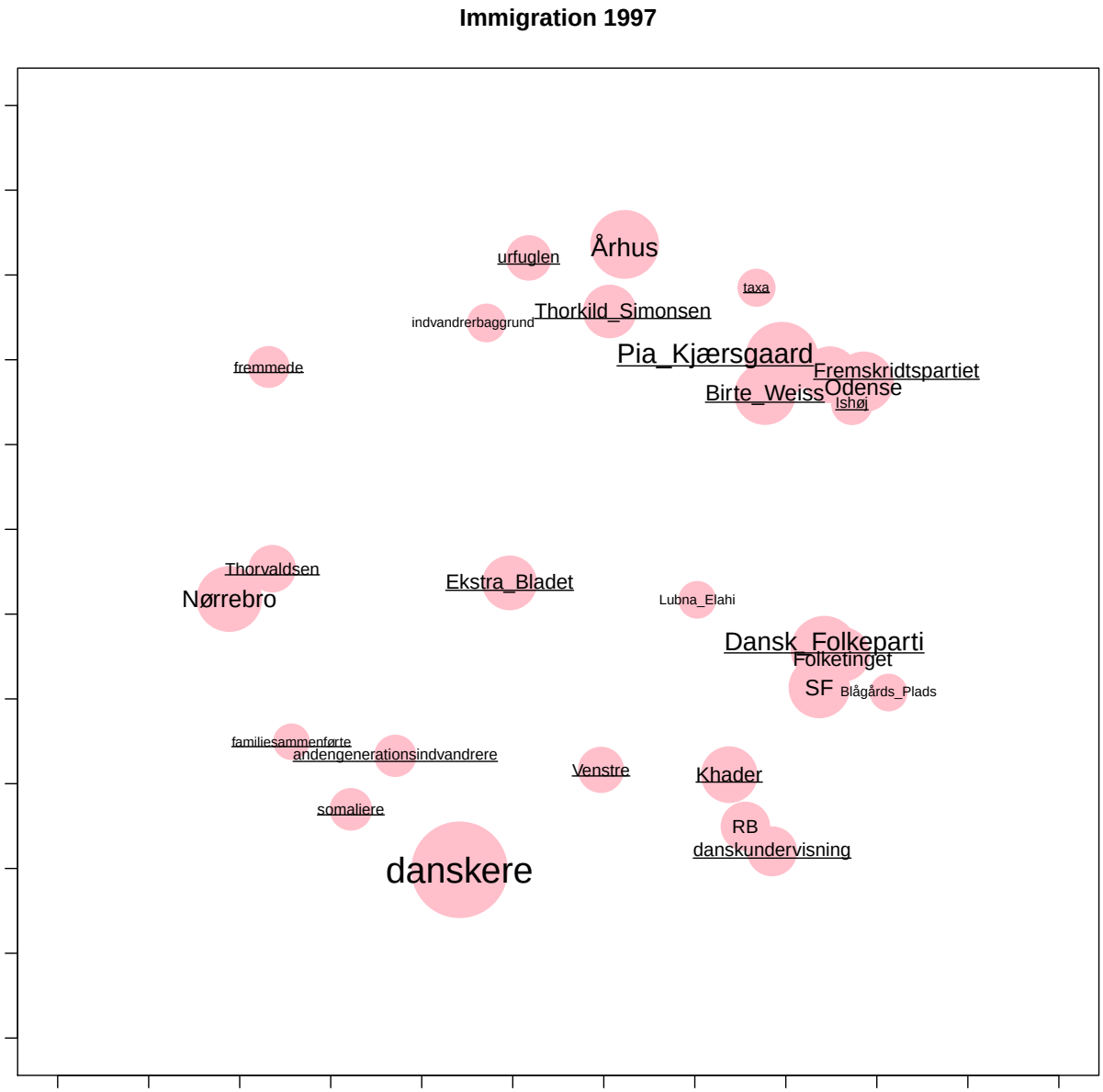

Figure 25: Visualization map for `Immigration` in 1997

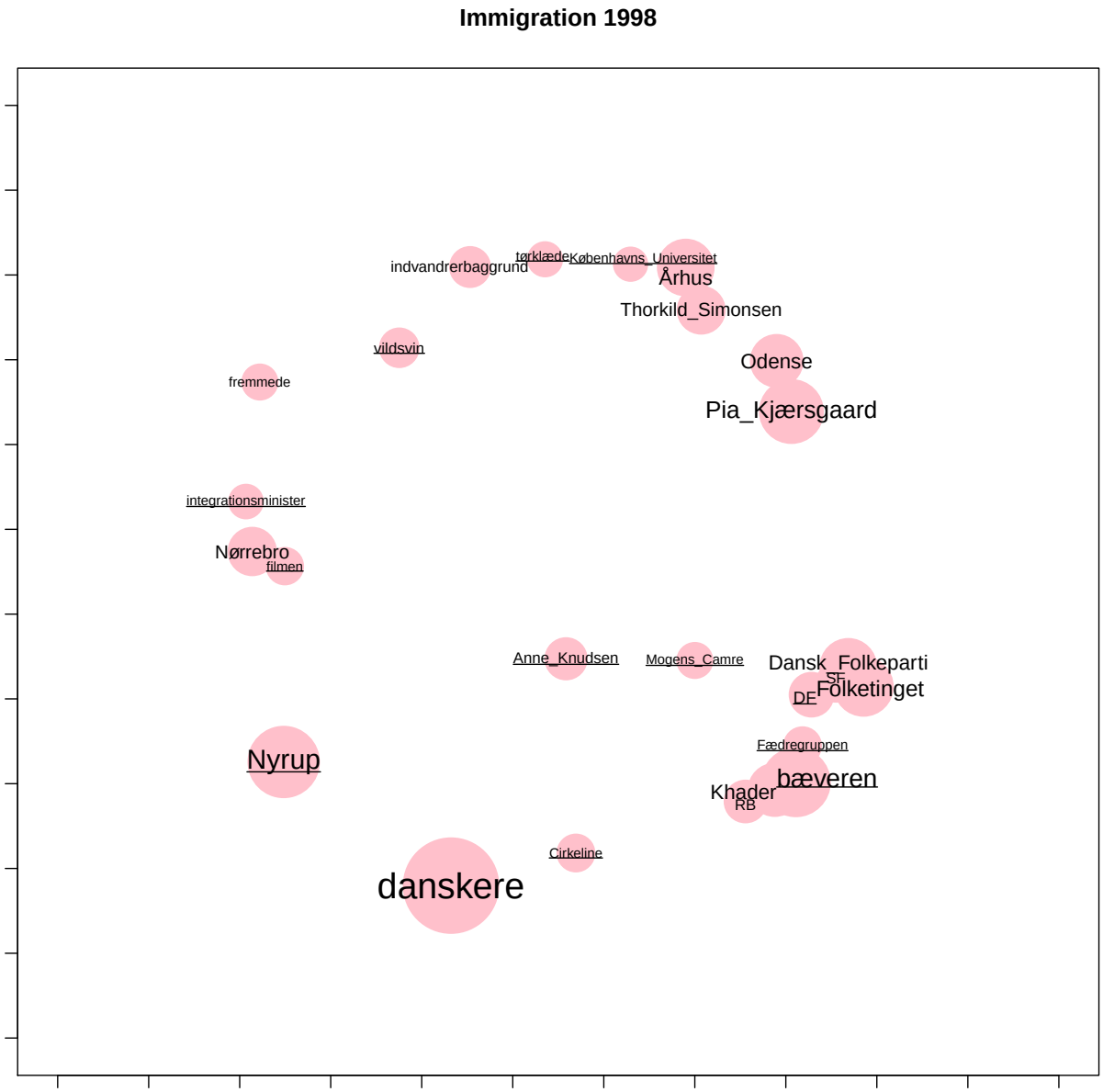

Figure 26: Visualization map for `Immigration` in 1998

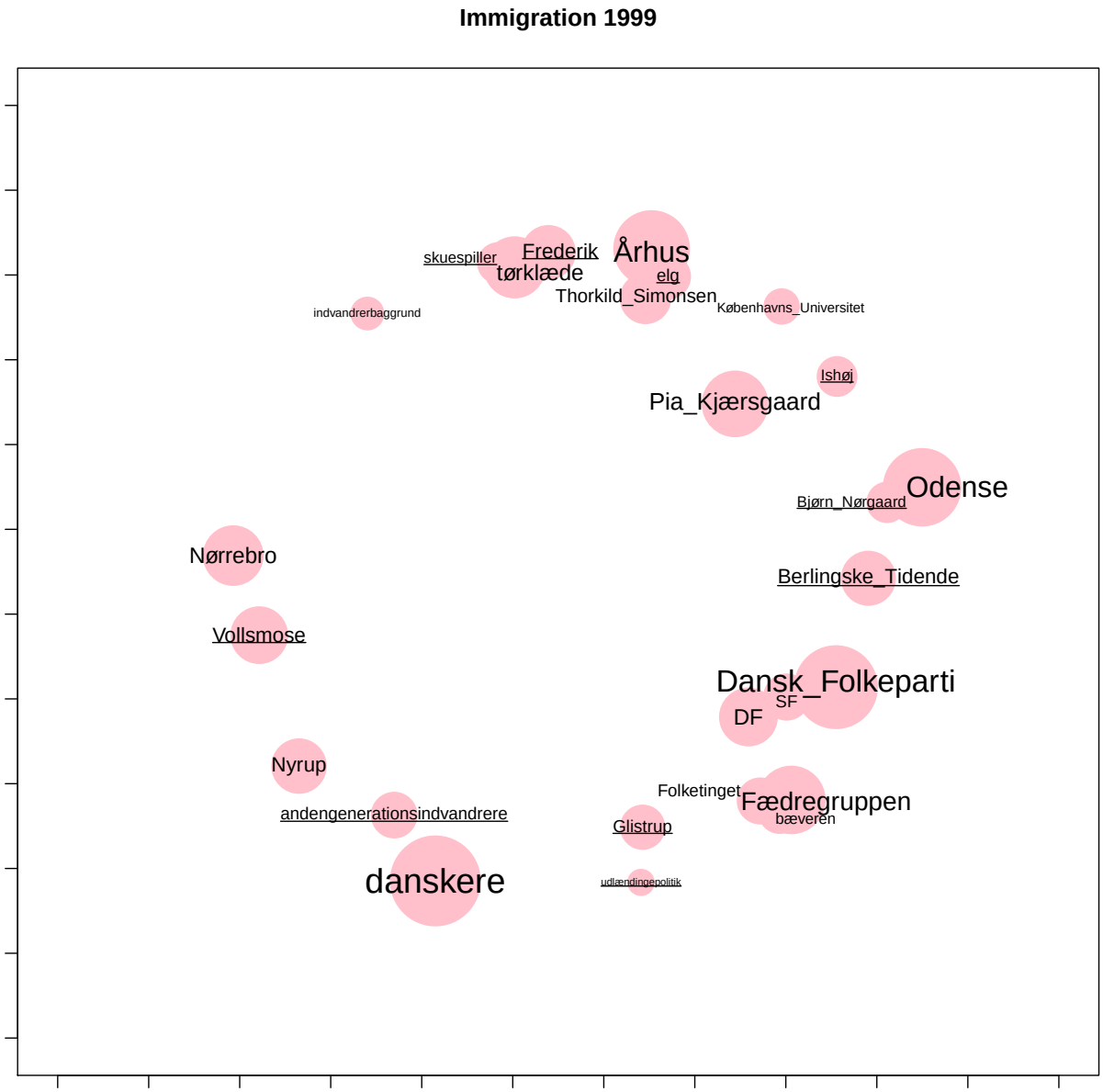

Figure 27: Visualization map for `Immigration` in 1999

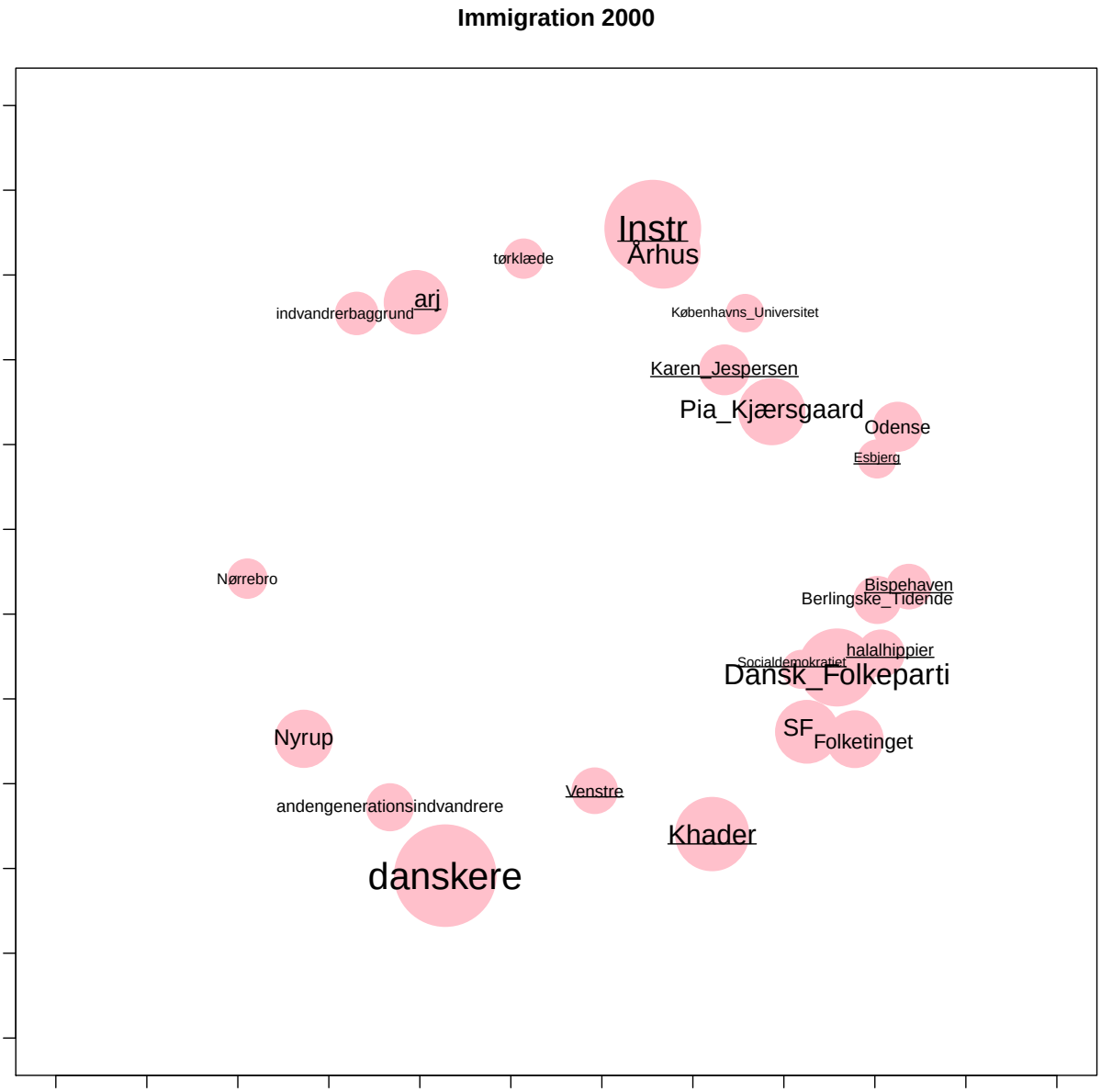

Figure 28: Visualization map for `Immigration` in 2000

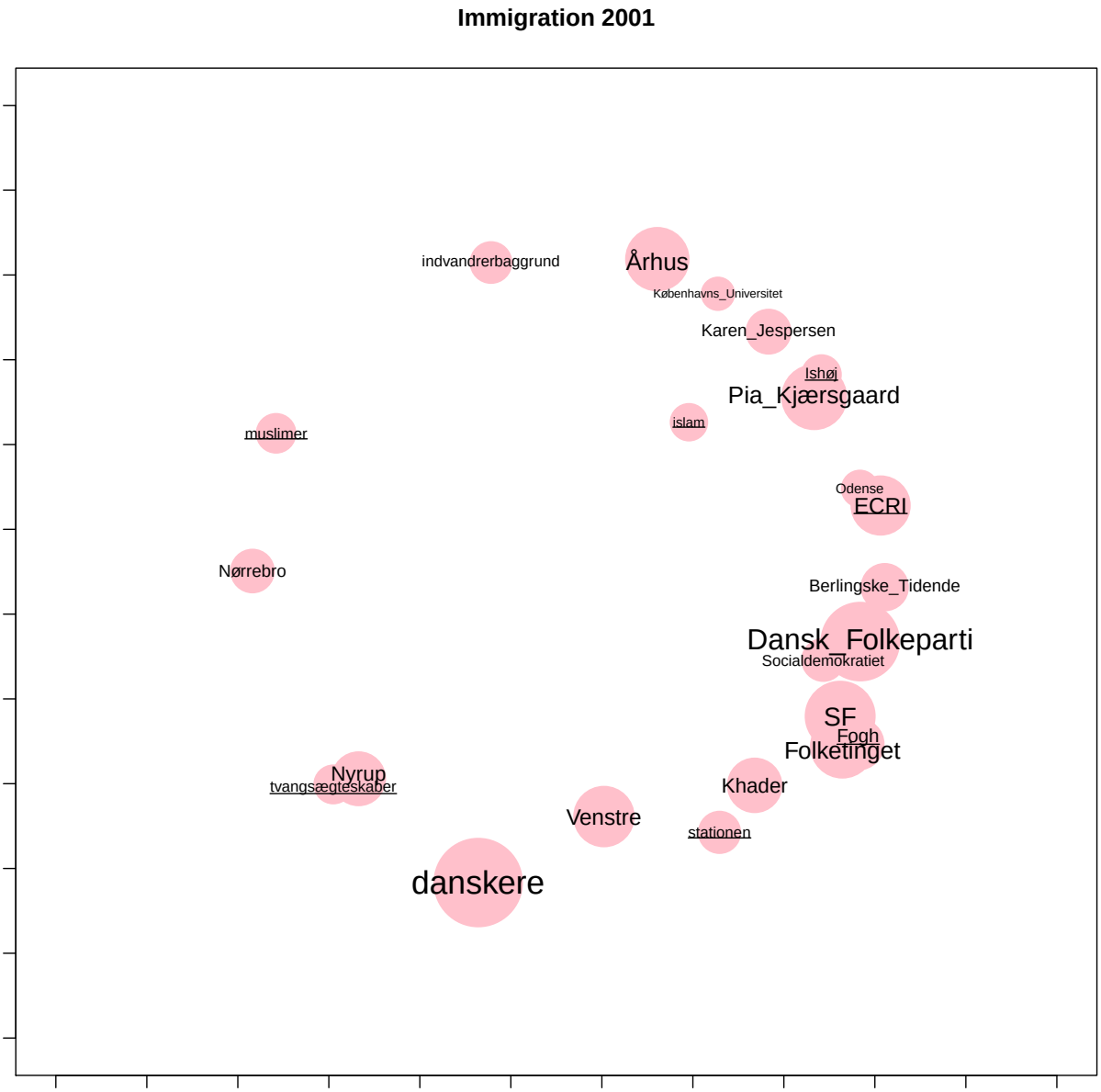

Figure 29: Visualization map for `Immigration` in 2001

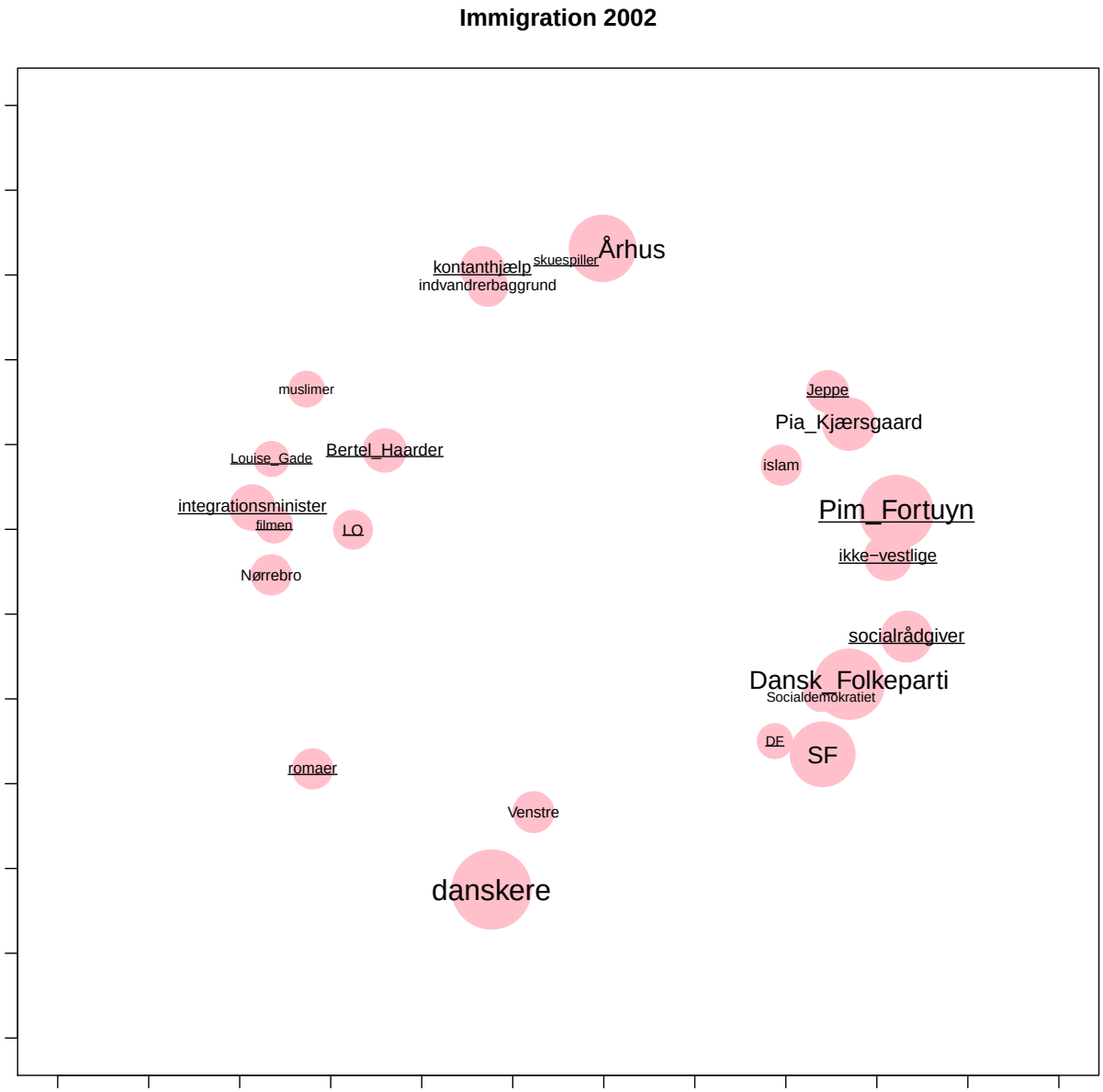

Figure 30: Visualization map for `Immigration` in 2002

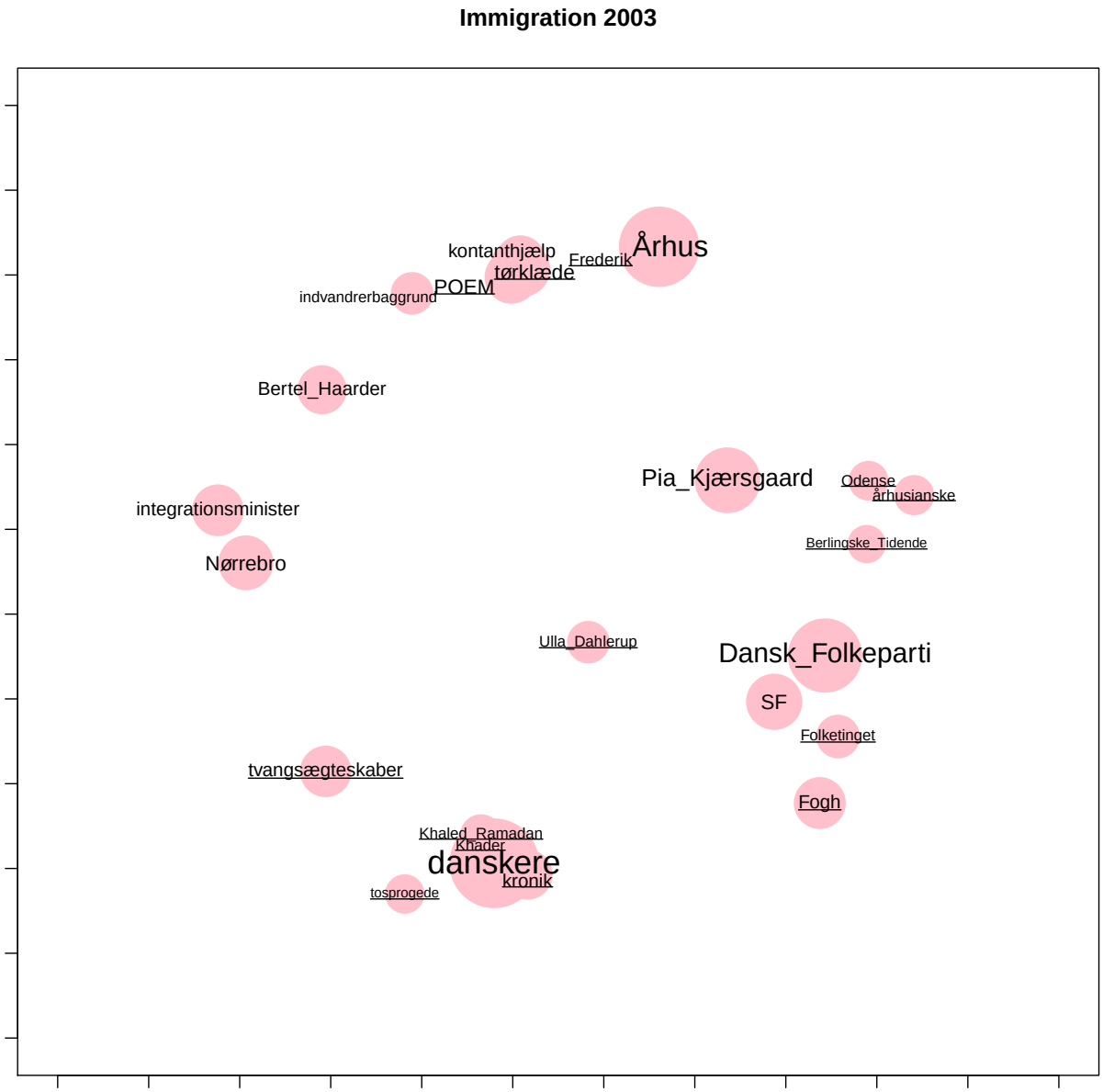

Figure 31: Visualization map for `Immigration` in 2003

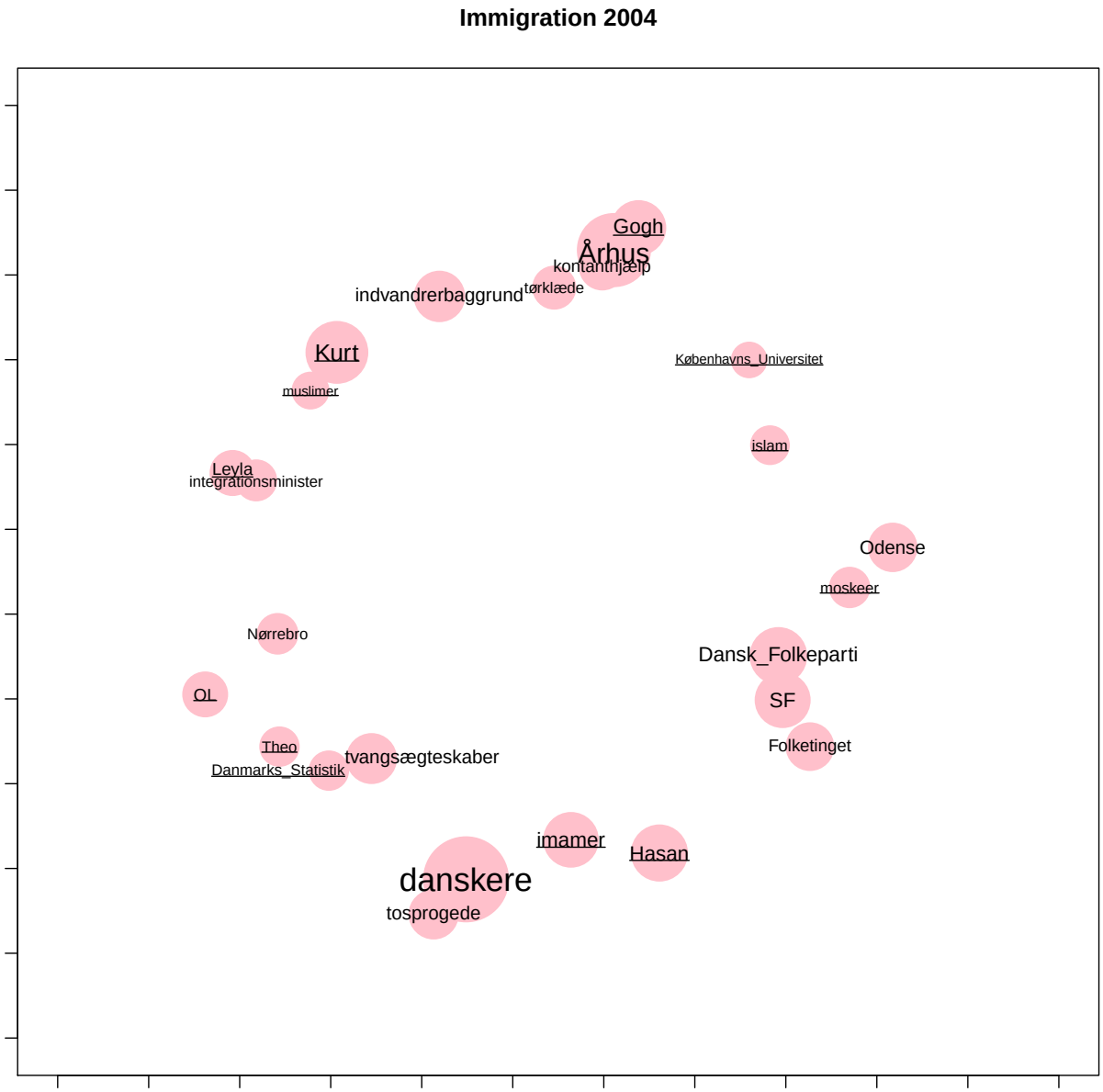

Figure 32: Visualization map for `Immigration` in 2004

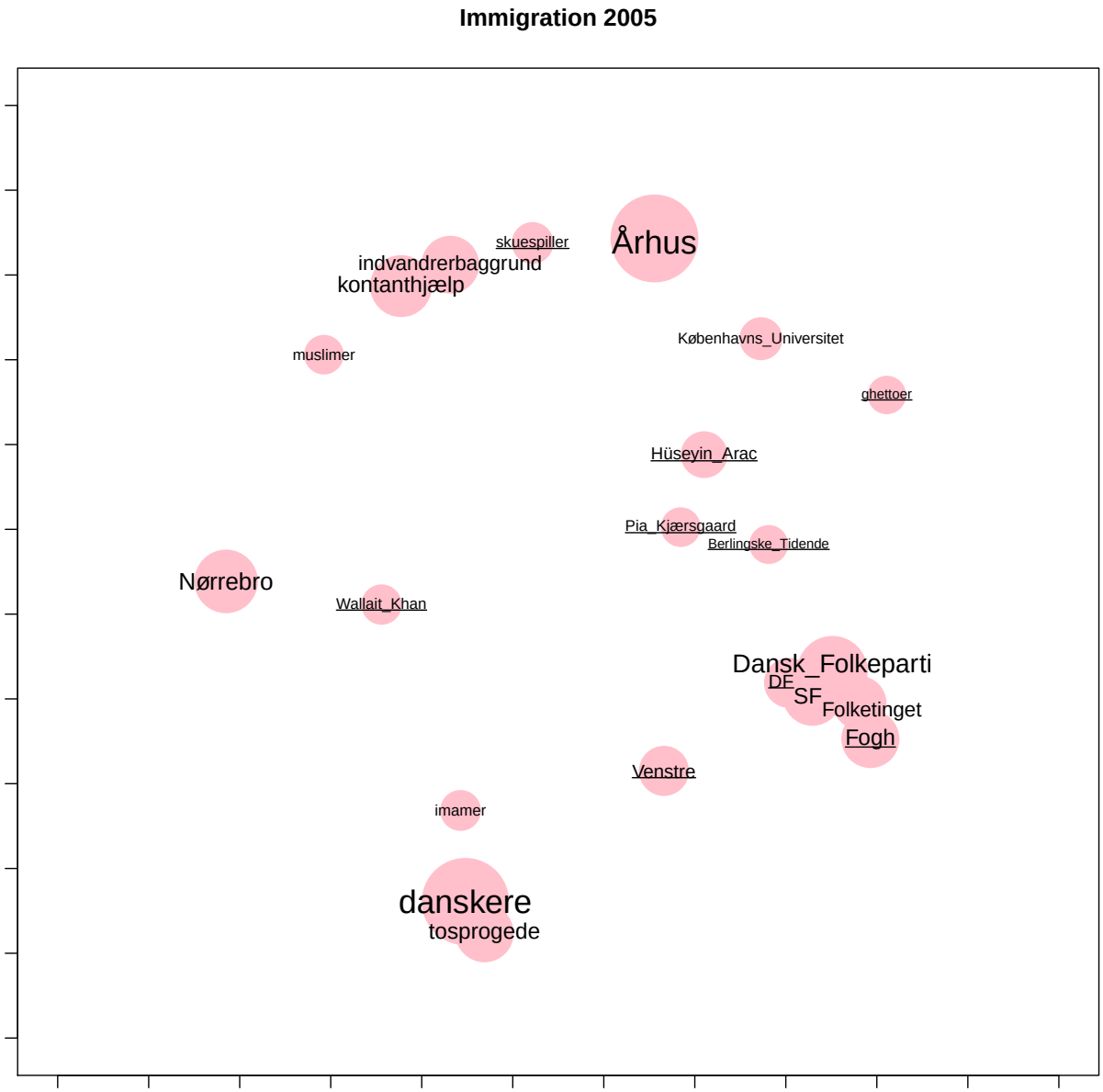

Figure 33: Visualization map for `Immigration` in 2005

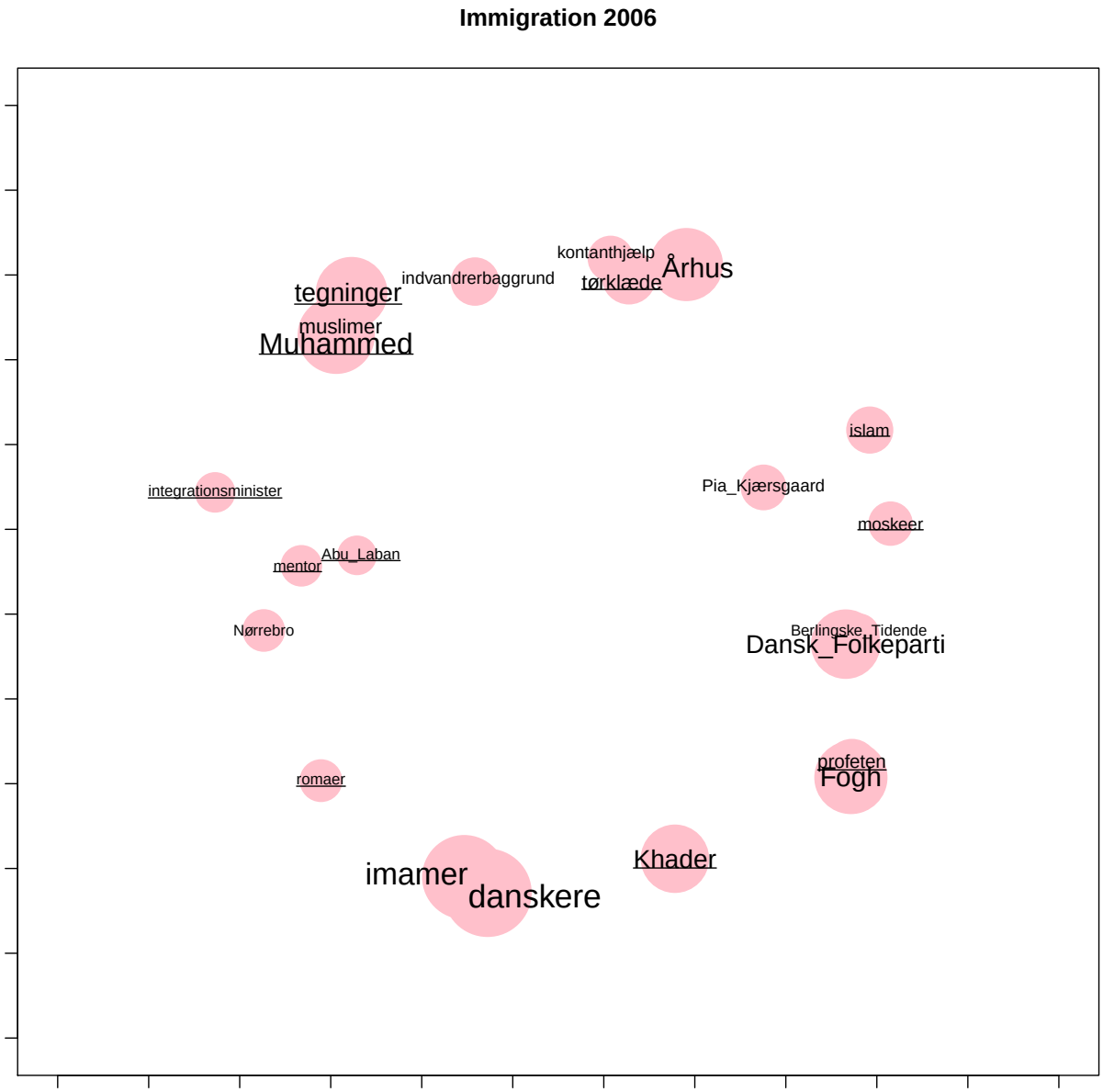

Figure 34: Visualization map for `Immigration` in 2006

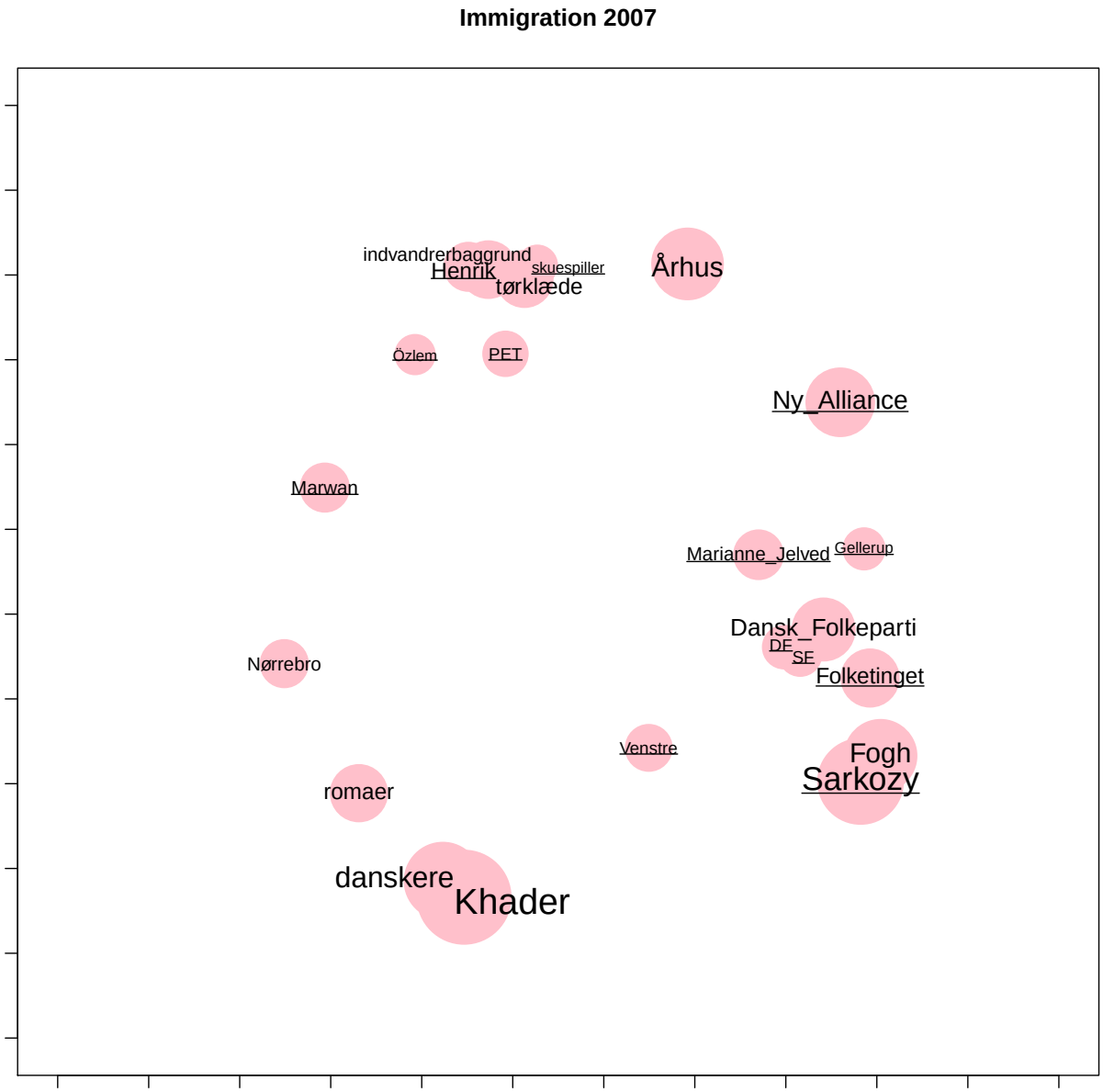

Figure 35: Visualization map for `Immigration` in 2007

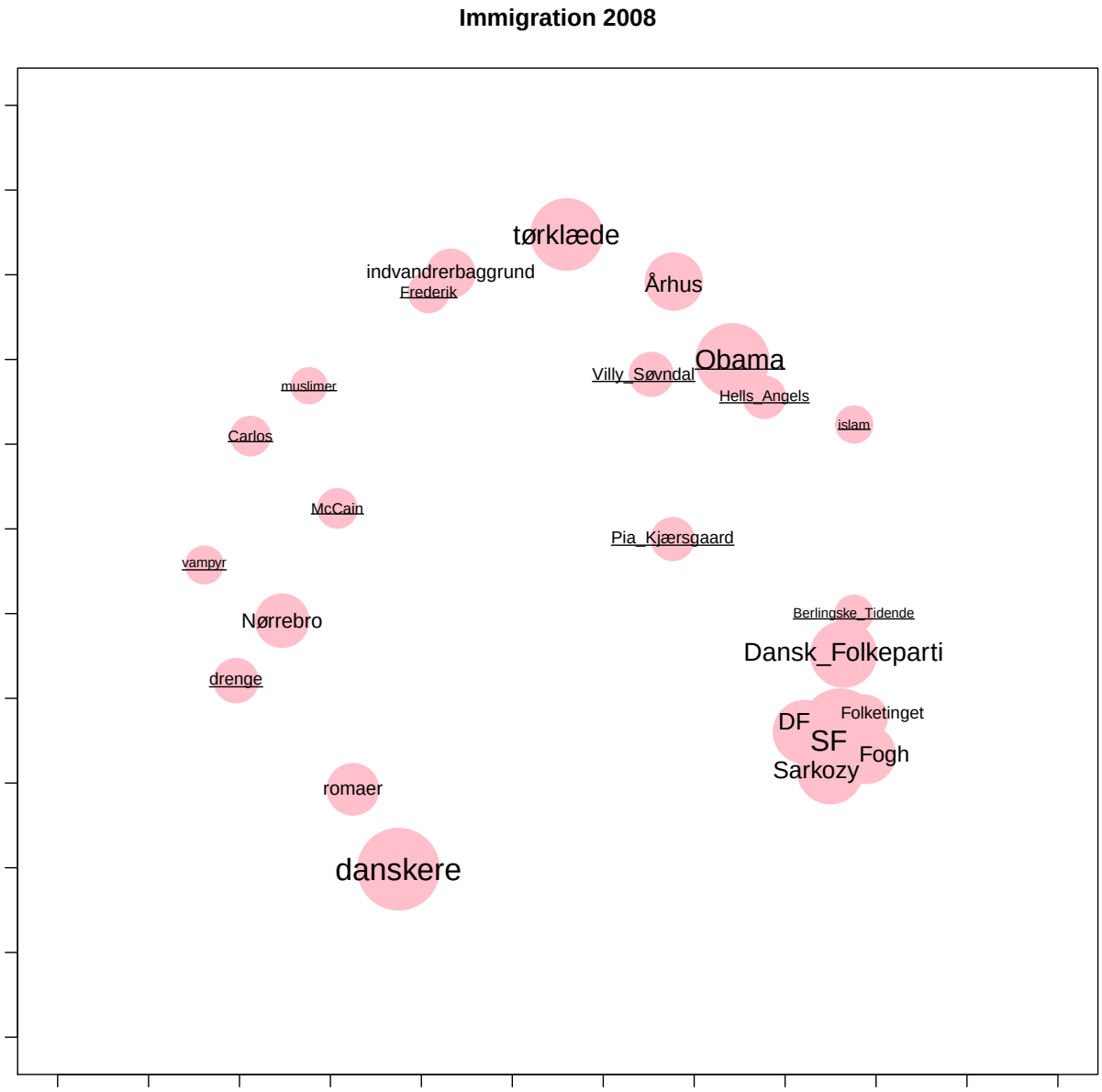

Figure 36: Visualization map for `Immigration` in 2008

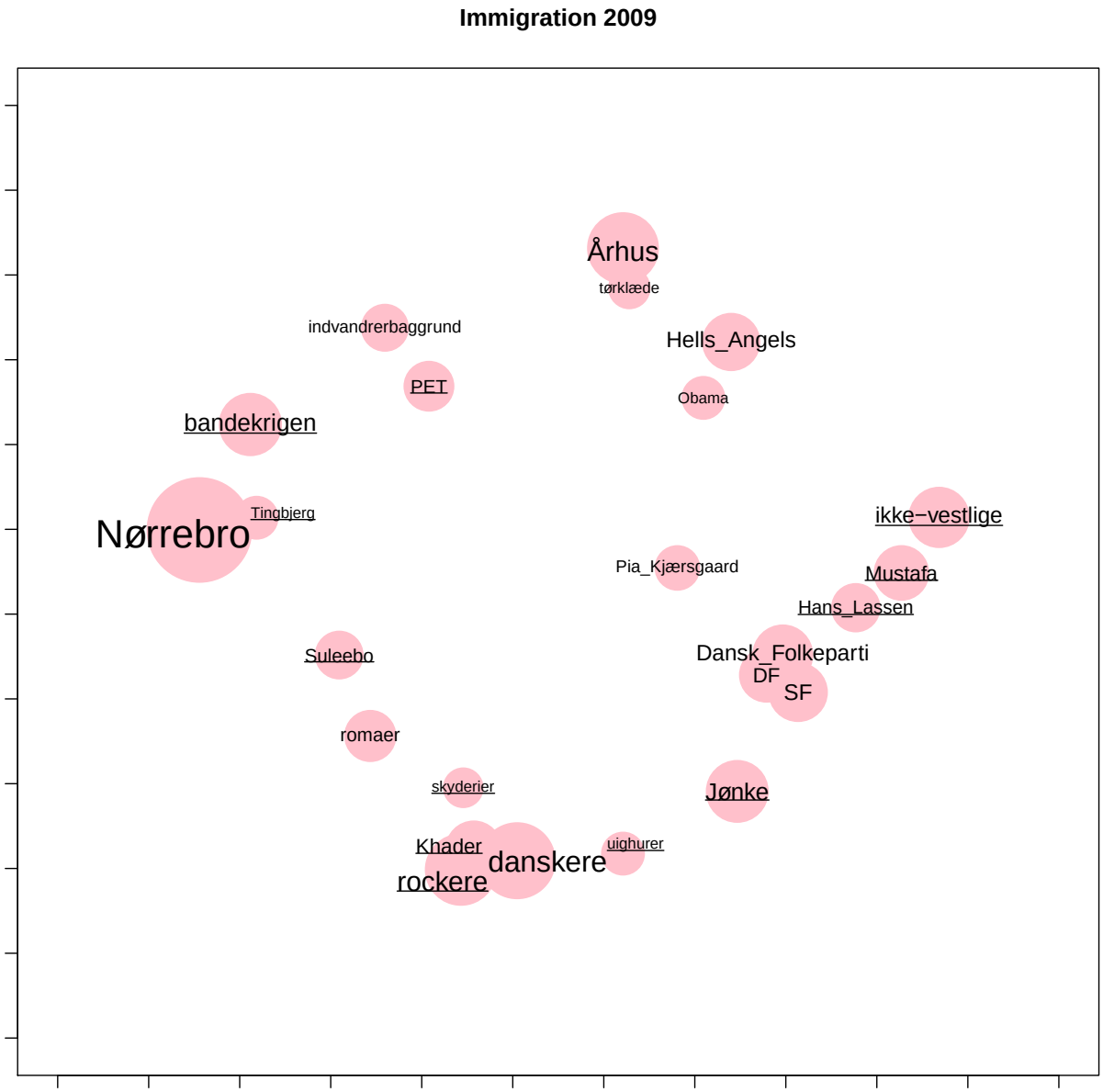

Figure 37: Visualization map for `Immigration` in 2009

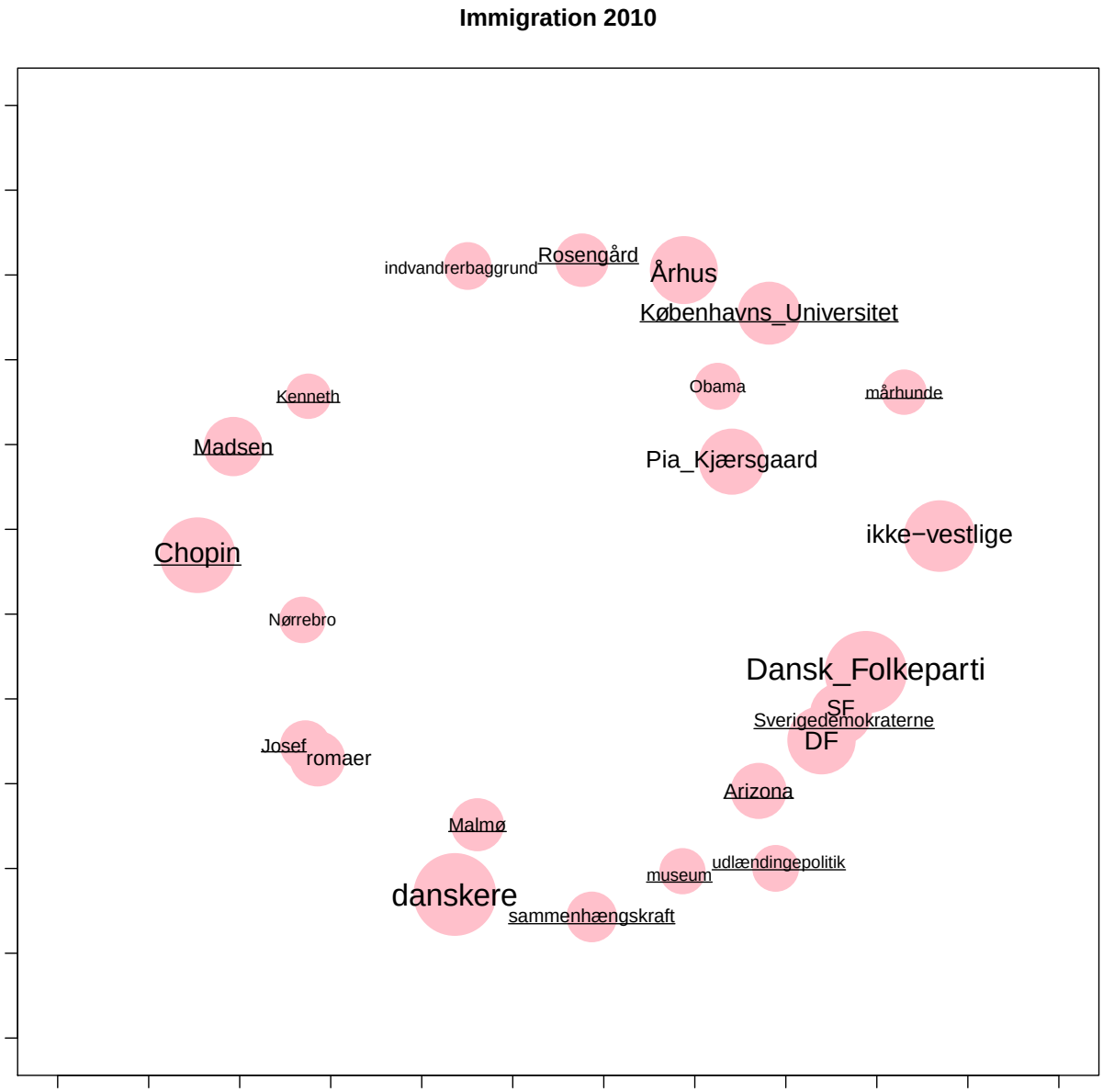

Figure 38: Visualization map for `Immigration` in 2010

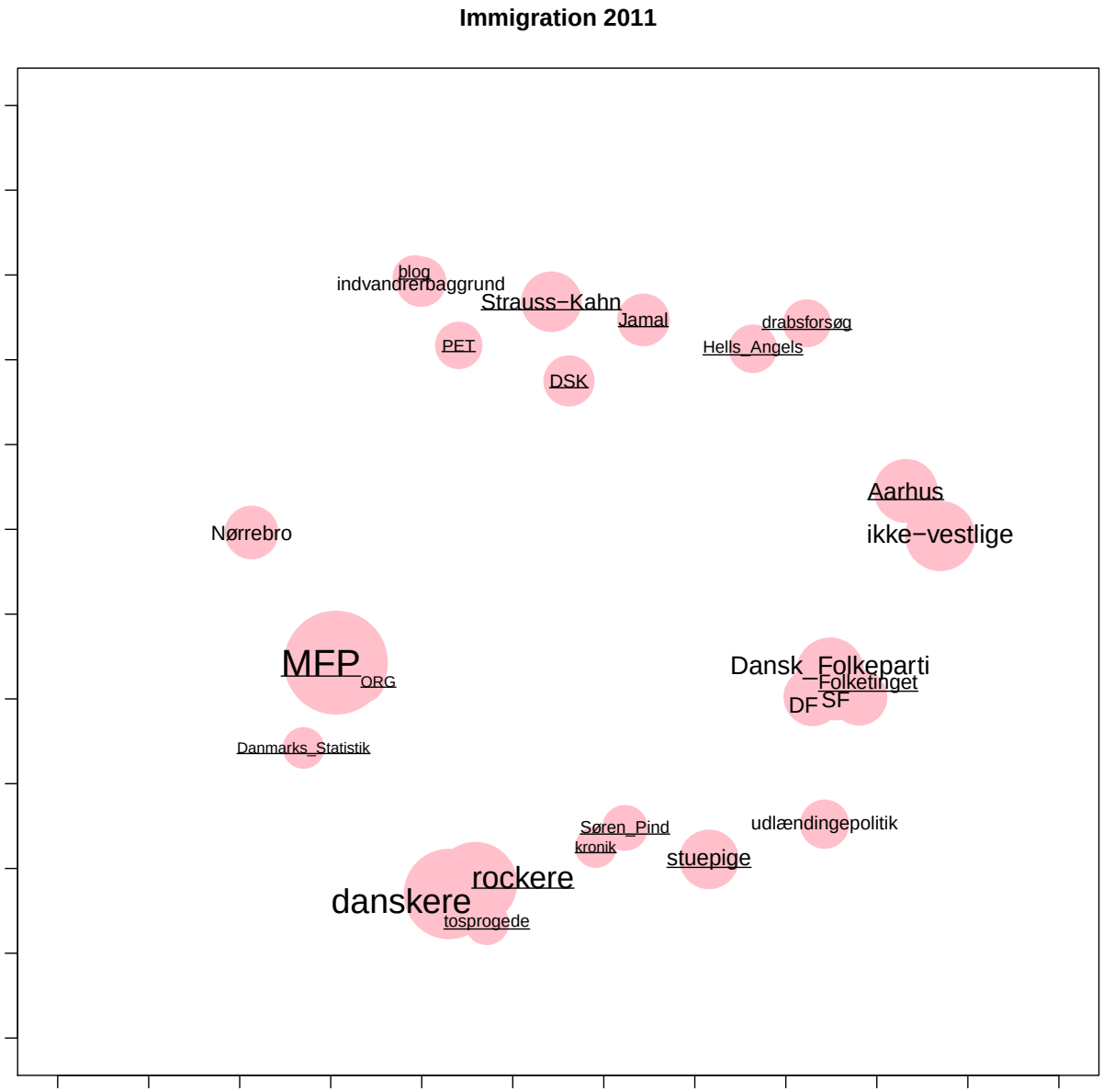

Figure 39: Visualization map for `Immigration` in 2011

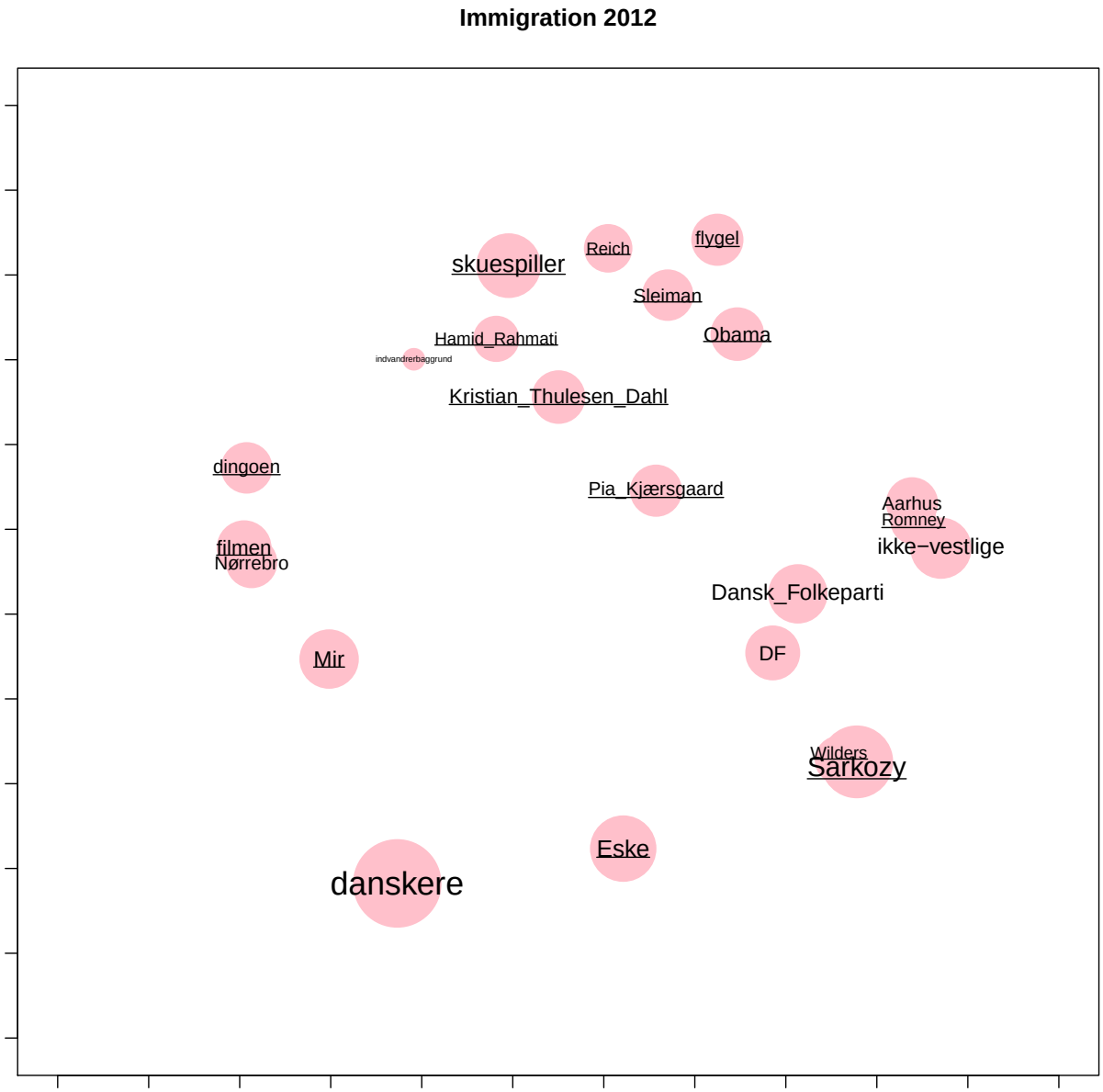

Figure 40: Visualization map for `Immigration` in 2012

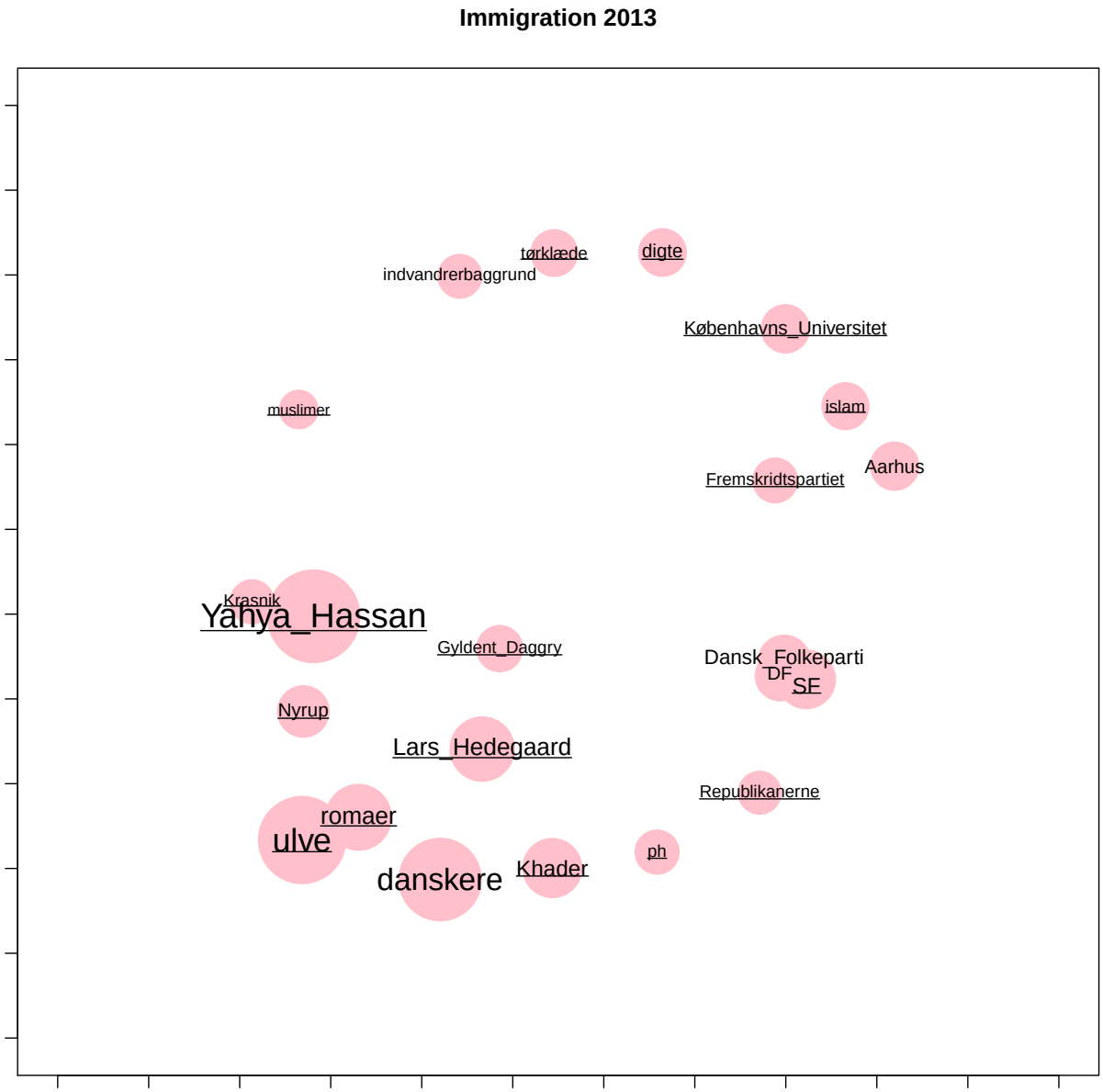

Figure 41: Visualization map for `Immigration` in 2013

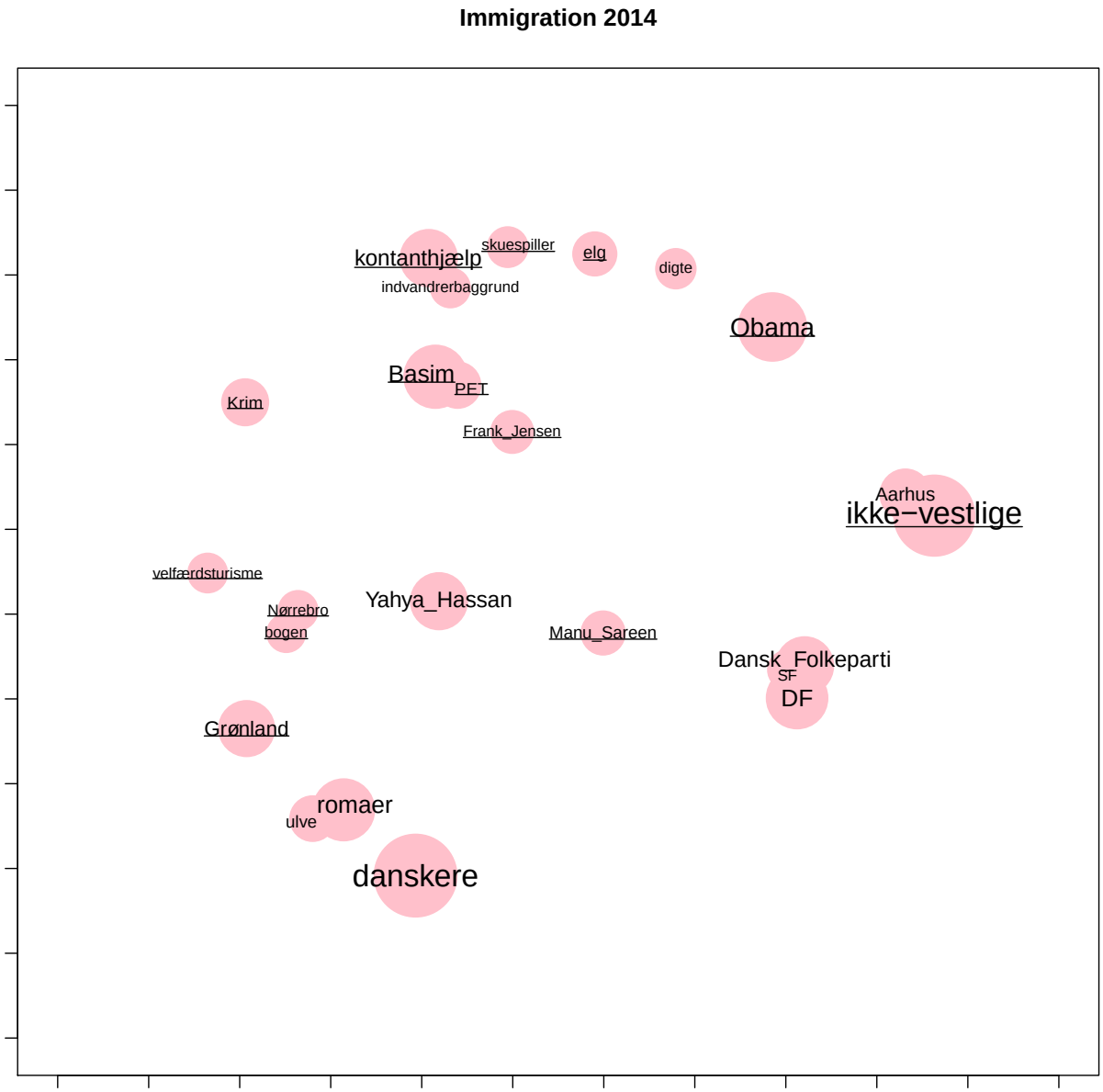

Figure 42: Visualization map for `Immigration` in 2014

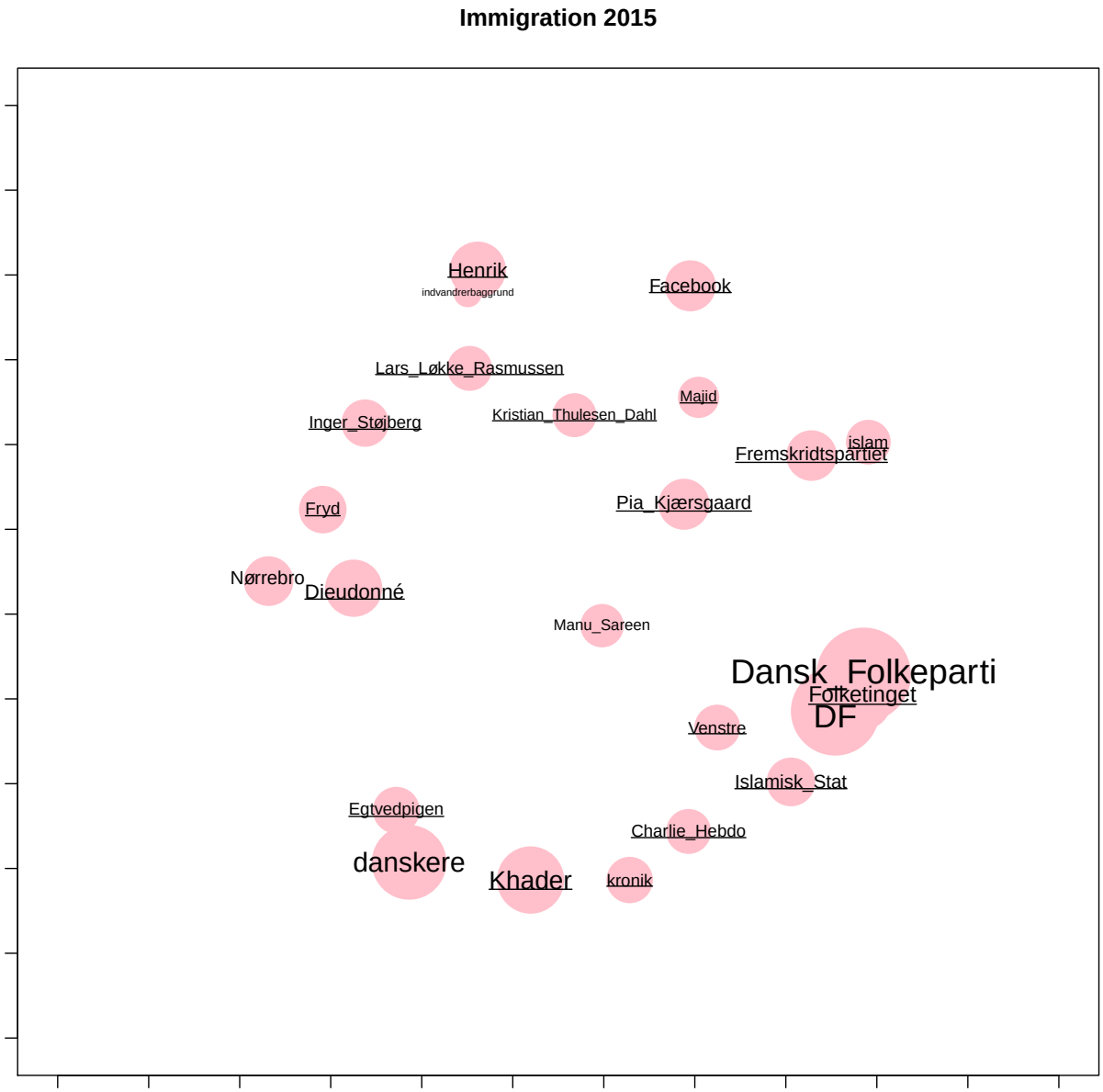

Figure 43: Visualization map for `Immigration` in 2015