Generative Data Models for Validation and Evaluation of Visualization Techniques

Christoph Schulz[1], Arlind Nocaj[2], Mennatallah El-Assady[2], Steffen Frey[1],
Marcel Hlawatsch[1], Michael Hund[2], Grzegorz Karch[1], Rudolf Netzel[1],
Christin Schätzle[2], Miriam Butt[2], Daniel A. Keim[2], Thomas Ertl[1],
Ulrik Brandes[2], and Daniel Weiskopf[1]

[1]University of Stuttgart, Germany [2]University of Konstanz, Germany

ABSTRACT

We argue that there is a need for substantially more research on the use of generative data models in the validation and evaluation of visualization techniques. For example, user studies will require the display of representative and unconfounded visual stimuli, while algorithms will need functional coverage and assessable benchmarks. However, data is often collected in a semi-automatic fashion or entirely hand-picked, which obscures the view of generality, impairs availability, and potentially violates privacy. There are some sub-domains of visualization that use synthetic data in the sense of generative data models, whereas others work with real-world-based data sets and simulations. Depending on the visualization domain, many generative data models are “side projects” as part of an ad-hoc validation of a techniques paper and thus neither reusable nor general-purpose. We review existing work on popular data collections and generative data models in visualization to discuss the opportunities and consequences for technique validation, evaluation, and experiment design. We distill handling and future directions, and discuss how we can engineer generative data models and how visualization research could benefit from more and better use of generative data models.

CCS Concepts

•General and reference → Validation; Evaluation;
•Human-centered computing → Visualization design and evaluation methods;

Keywords

Generative data models; visualization; validation; testing; evaluation; benchmarks; user study design

1. INTRODUCTION

In contrast to manual data collection and production, generative data models leverage computers to generate data, distinguish instances among the ones generated, based on common properties and thus define what is considered *valid* and *representative* data. Generative models allow us to address difficult challenges when designing, testing, benchmarking, and assessing visualization techniques in general: How can we extract and design around data-inherent constraints? How can we validate an implementation? How do we reason about the visual response of a system to data? How can we measure technical, visual, and perceptual scalability? How do users interact with a visualization? How can we verify that information is understood well? As researchers, we need to conduct empirical experiments to find answers to these questions [102].

Others have reviewed and discussed the practice of evaluating visualization systems [53, 65], albeit focusing on the processes of designing, performing, and evaluating studies. We are convinced that data itself should receive more attention to improve research quality by eliminating confounding variables for testing a hypothesis and obtaining a link between a data model's parameter space (input) and the visualization result (output) for statistical analysis. In this way, we will be in a better position to prove correctness and completeness of a technique, i.e., without introducing statistical bias through samples with questionable representativeness. A trend toward increase in both data size and complexity [2] could prove as a bottleneck in collecting data for user studies, lab studies, performance studies, and testing. Abstract and formal descriptions are more compact, easier to transfer, and thus leverage *availability* for experiments, while usually also being more tangible and thus beneficial for analysis. For *reliable* empirical research results, we need to be able to conduct controlled experiments and isolate what we want to measure. Yet, input data during our experiments is often cluttered with confounding variables. We are convinced that generative models will help us produce data with different characteristics to allow comprehensive data collection and a thorough evaluation of user studies and studies of technical performance alike. Finally, ethical and legal issues are often connected to data collection, in particular, *privacy* and *property*. For example, institutions or companies are reluctant to publish medical, social, or personal data. The current situation of data collection often hinders development and research of new visualization techniques, while generative models are scattered across the visualization community with many of them being “side projects” of techniques papers, and thus not reusable, let alone general-purpose.

With this combination of position and survey paper, we

BELIV '16, October 24 2016, Baltimore, MD, USA

ISBN 978-1-4503-4818-8/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2993901.2993907

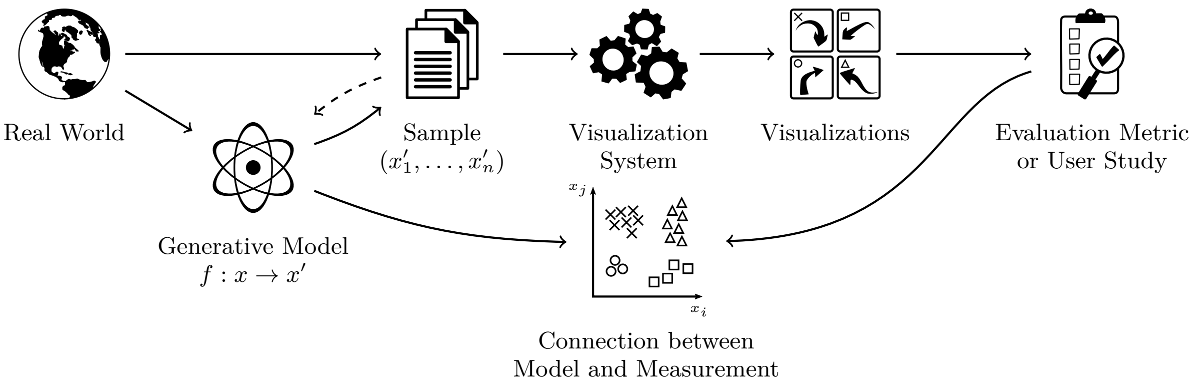

Figure 1: Workflow for validation and evaluation of visualization techniques using generative data models.

want to focus on generative data models as part of visualization research. The (larger) part of the paper reviews relevant examples of such data models to understand the state of the art across the main subfields of visualization. In this way, we arrive at a substantiated perspective on interesting future directions that will lead to a better and more frequent use of these data models.

2. OVERVIEW

A *generative data model* is a simplified and idealized understanding of an observation, capable of generating data values. This definition is not to be confused with *generative models* from machine learning, having a far more narrow definition of what a model is. A related term is *procedural content generation*, known from computer graphics or games. The techniques used are similar, but the goals are different: the computer-graphics term emphasizes consumability, whereas the visualization community's term emphasizes hypothesis testing. Figure 1 shows our abstraction of the workflow for experiments in visualization research that work with generative models. A generative model is derived from a theoretical or real-world observation to generate samples, as opposed to measured samples. The dashed line is for hybrid generative models. The visualization system is put into the loop for validation or evaluation purposes.

The visualization community aims for controlled experiments as an important element of empirical research. We restrict our discussion to the following relevant requirements in terms of data: *validity, representativeness, availability*, and *privacy*.

Opportunistic data collections are inherently valid, given the collection method is valid, but representativeness might be questionable. Collected samples may be very large and hence technically complicated to store and transfer, hampering availability. Furthermore, it is much too easy to apply bad anonymization, impinging on privacy issues with severe personal and legal consequences in the real world.

Generative models, in contrast, are an abstract description, and thus do not contain any private information. Yet, they are easy to store and transport, but their validity is always in question because models are a picture of the world based on its author's beliefs and values. From our point of view, validity depends on the target experiment requirements. In practice, a good generative model does not necessarily mimic real world observations in their entirety, but just the aspects to be tested. Because a generative model is also a formal description relating input parameters to output data, representativeness is well defined, which allows us to connect model and measurement, enabling exhaustive analysis techniques.

Table 1: Tagging scheme for generative data models.

Tag	**Meaning**
Parameter Extraction	
rewriting systems	replace subterms of a formula with other terms
(local) search	means moving from solution to solution in the space of candidate solutions by applying changes
imitation	is based on an observed process
Parameter Type	
local	establishes a local property
global	establishes a global property
forward problem	parameter manipulation happens in the model's domain
inverse problem	parameter manipulation happens in the model's codomain
Technique	
measured	defines a magnitude of some attribute of an object relative to some unit of measurement
user-defined	requires user input to determine an analytical or numerical solution
hybrid	diversity is generated based on top of a non-generative data model
random	model inherently contains a random variable
deterministic	model contains no randomness
closed-form	an expression that can be evaluated using a finite number of operations
sequence	model generates a sequence of data, e.g., data ordered by time
emulation	imitation of the end result of a real-world process behavior, i.e., that does not approximate intermediate steps
simulation	imitation of the operation of a real-world process or system, i.e., that approximates intermediate steps

Although using generative models for visualization experiments sounds easy and tempting, there are many technical obstacles to deal with in practice. How can we synthesize data with certain desired properties artificially, while the generated data exhibits the relevant characteristics of real data to guarantee the external validity? How do we relate model parameters to visualization results? These are difficult questions from a scientific and engineering perspective.

Based on our findings from the survey, we present a tagging scheme for generative models, organized into three categories, shown in Table 1. The first category is about high-level parameter extraction. Presumably all model parameters can be obtained using rewriting (e.g. L-systems), search (e.g. heuristics), or imitation (e.g. simulations), each with different trade-offs. The second category is about how to think about model parameters: whether a parameter establishes a global property (i.e. all) or a local property (i.e. there is) of a dataset, and where interaction happens (i.e. forward or inverse problem). The third category is about techniques used for generating data.

3. LITERATURE SURVEY

This section surveys of the state of art of generative data models related to visualization. The survey is organized according to the classification of visualization methods by Tory and Möller [112]: typical techniques from scientific visualization (scalar, vector, and tensor field visualization) and information visualization (multi-dimensional data and network or graph visualization). The organization has been extended to incorporate recent trends within the visualization community, in particular, trajectory and text visualization. Most real-world visualization and visual analytics systems are not restricted to a single visualization technique from our structure but can deal with various combinations of data types and techniques. We do not explicitly cover such *hybrid* combinations in our survey; however, many of the single-technique data models will carry over to such hybrid approaches and therefore serve as a good basis.

To conduct our survey, we gathered experts from different visualization domains to summarize the state of data collections and generative models within their domains. In an iterative process, we refined the tagging scheme (Section 2) to reflect the main aspects of the paper collection and to update the tagging assignments to the papers. This process is similar to previous approaches to literature surveys [8]. The following discussion of the literature explicitly includes the tags *in italics* for techniques employed in generative data models; the other tags are mostly used internally.

Our approach is systematic in the sense that it covers the relevant areas of visualization techniques. However, it is not meant to be a comprehensive survey of papers in a certain field. We have been facing the problem that there are few to no relevant papers in the visualization literature for some of the visualization techniques. Therefore, we include papers from other fields that have some connection to generative models for visualization. Due to this breadth of possible journals or conferences of interest, we chose to include the expertise of the visualization experts to select representative sets of papers.

The following review will also show that many subdisciplines of visualization lack an extensive use of generative data models. For a complete picture, we include a few pointers to *measured* data that do not come with any synthesization method, i.e., which are not generative in nature.

3.1 Scalar Field Visualization

Scalar fields assign a single (scalar) data value to each point of a dataset. For visualization, mostly 2D and 3D spatial domains are considered, and in many cases they are time-dependent as well. Depending on the area of application, different data with different characteristics are considered representative, and there is a variety of available datasets for different domains, such as the CT Dataset Archive[1], OASIS[2], or DWD GDS[3]. For the evaluation of generic volume visualization techniques, it is common practice to use a number of datasets from repositories such as volvis.org[4].

An important class of scalar field data comes from *measurements* obtained via scanners, most commonly for medical but also for industrial applications, like material testing. For example, computed tomography (CT) scanners reconstruct a volume for several X-ray projections [115]; magnetic resonance imaging (MRI) scanners use magnetic fields and radio frequency pulses [105]. Depending on the technology used, data can be static or time-dependent. Sometimes, computer simulations are employed to model the scanning procedure, both for the purpose of research and education [5]. However, most of the scanning-based datasets are just measured and cannot be used as a basis for controllable data generation.

Another common source of scalar data are *simulations* employing physically-based models[5]. They typically output several quantities that can be scalar volumes, but e.g., also vector fields (Section 3.2). Apart from the large field of flow simulations, there are other types like the estimation of a fire danger index for geographic regions based on vegetation, topology information, and weather data [20]. Also, note that new volume data can be produced using these simulations as input, e.g., via the extraction of similarity information and temporal clustering [39]. Although simulation-based data models allow us to generate new datasets, e.g., by re-running the simulation with different parameter settings, these simulations are not designed to serve as controllable data sources for empirical studies.

The third class of datasets are derived from *closed-form* representations. The classic example is Buckminsterfullerene—the so-called bucky ball—the spherical C_{60} fullerene molecule. Other examples include the hydrogen atomic wave function, equations from quantum mechanics, and 3D fractals. Their representation is resolution-independent, differing from the typically (grid-based) data by measurement and simulation, and it is therefore the perquisite to studying certain characteristics of visualization techniques [38]. Analytic representations are closer to the needs of generative models, but many of these representations have no steering parameters or the parameters cannot be easily used to control the data generation for empirical studies.

Apart from previously discussed scientific data, there are related data models from computer graphics. For example, modeling and rendering tools like Renderman[6] are used to generate clouds. Solid texture synthesis algorithms automatically generate volumes from a set of 2D example images [90]. Other approaches restrict synthesis to a subset of the voxels of interest, such as a single surface layer [28]. There are procedural models for which the user specifies material information at certain layer depths for given 3D surfaces using a scripting language [24]. Similarly, there is a sketch-based system to design volume data from scratch [86]: the user splits a surface model and paints brush strokes to volumet-

[1] http://www.sci.utah.edu/cibc-software/

[2] http://oasis-brains.org/

[3] http://www.dwd.de/DE/leistungen/gds/gds.html

[4] http://volvis.org/

[5] http://vis.cs.ucdavis.edu/VisFiles/pages/dataset.php?ds=1

[6] https://renderman.pixar.com/view/tgt__volumes

rically fill the 3D space. Finally, different types of data may also be combined, e.g., simulated data can be used as input for manual modeling, or measured data may be altered procedurally. In short, the computer graphics approaches provide some examples of *user-defined, simulation*-oriented, *measured*, and *closed-form* generative models, which could inspire visualization research. However, there is no model that would directly lead to generative models for empirical tests of visualization.

Scalar data exhibits different properties that can be used to test certain characteristics of visualization techniques. They may be altered for evaluation via a variety different approaches outlined below, some of which are 3D extensions of image processing techniques[7]. Scalar fields may be given via different representations, and resampling and conversion between them can be done to evaluate both the accuracy and performance of visualization techniques [37]. The stability under noise is an important property for many types of data, and it may be applied additionally on the data for evaluating robustness [39]. Finally, many techniques differ for static versus time-dependent representations.

3.2 Vector Field Visualization

Vector field visualization techniques are often used for the analysis of fluid flow. The structure and complexity of the vector field depend on the type of visualization technique to be tested. A simple *closed-form* flow may be sufficient to measure the extent of numerical diffusion introduced by a texture advection method [66]. In contrast, we need more complex datasets with time-dependency, stretching, and folding behavior to assess visualization techniques based on Lagrangian coherent structures (LCS) [46]. Such data can also serve as a validation for integral surfaces where care must be taken to adapt the meshing to the stretching and thinning regions of the flow. Moreover, for flow visualization techniques (as opposed to general vector visualization techniques, e.g., for magnetic fields), datasets should be solutions of the Navier-Stokes equations. This is important in the physical validation of the visualization techniques. For topology visualization [67], the existence of sinks and sources is desired to test the robustness of detection and tracking methods for critical points.

One class of complex vector fields comes from *measurements* of experimental data. For example, particle image velocimetry (PIV) allows one to measure the velocity vector field of a fluid flow. PIV data were used to assess texture-based [13] and topology-based [85] uncertainty visualization. A common problem with the experimental datasets is that they usually require some pre-processing to, e.g., filter out the noise. Moreover, such data has typically limited spatial and temporal resolution, and scaling the datasets is not trivial due to highly non-linear flow behavior.

Another class of vector fields is given by *user-defined* data. For example, image-based flow visualization [118] comes with a demonstration program in which the user can hand-pick the locations of critical points in the vector field[8]; also, one can assign velocity to those elements to achieve time-dependency.

Laidlaw et al. [64] employed *random* vector field generation: they selected randomly nine positions within a unit square where the vectors were again chosen randomly. From these points, the vector field on the whole domain was generated by interpolation. Liu et al. [73] adopted a different approach, in that they defined the critical points explicitly and generated structurally different vector fields with similar topological complexity. Such user-defined datasets allow for optimization by their *local* properties, i.e., critical points.

Simple *closed-form* models can be used to assess the quality of texture advection methods. For instance, rotational flow can be used to demonstrate visual artifacts in dye advection [58, 71]. LCS can be investigated with an analytical model of a double-gyre flow, whose parameters allow for both steady and unsteady flow [107]. Spherical vortex models [100] can be extended by swirl and tilt to obtain phenomena that occur in nature [89]; although the modified version is no longer a solution to Navier-Stokes equations, it provides flow characteristics that help test certain visualization methods. An often used data model is the "tornado" dataset[9] that is defined by an analytic expression and lends itself to testing interactive texture-based visualization methods [21, 123].

There are many works in vector field visualization that demonstrate the techniques on "standard" datasets that are numerical solutions of Navier-Stokes equations. However, each paper uses its own *simulation* result with custom parameters, so quantitative method comparison is difficult. Examples of such datasets are the flow around a cylinder with Von Kármán vortex street[10] [119], Taylor-Couette flow [7], and delta wing [40].

3.3 Tensor Field Visualization

Tensor fields are used to describe complex phenomena that cannot be sufficiently represented with scalar or vector fields, e.g., mechanical stress or diffusion processes. Visualization research focuses mainly on symmetric second-order tensors and two application areas: Engineering applications use tensor fields, e.g., to analyze mechanical stress or fluid flow. In medical applications, diffusion tensor imaging (DTI), a special variant of magnetic resonance imaging (MRI), is used to measure the diffusion of water molecules.

There are some repositories for measured DTI data, e.g., data from the Human Brain Project[11], from the Human Connectome Project[12], from the Alzheimer's Disease Neuroimaging Initiative[13]. However, there seems to be no long tradition of common benchmark data in this field. Therefore, the MICCAI DTI Challenge[14] [95] was established. There seem to be no open repositories for other types of tensor fields.

The visualization and analysis of diffusion tensor fields is an important topic in medical research [69] and the respective methods are usually evaluated with *measured* DTI data from real patients [95]. To obtain DTI scans with specific properties, phantoms are used. Phantoms are physical objects with specific properties, e.g., they consist of capillaries to emulate fiber structures [114, 126]. Typically, such phantoms are handmade and generative models are not involved. However, this might change since experiments with 3D printed phantoms are performed [82]. There are also approaches that

[7] https://www.wolfram.com/mathematica/new-in-9/3d-volumetric-image-processing/

[8] http://www.win.tue.nl/~vanwijk/ibfv/

[9] http://web.cse.ohio-state.edu/~crawfis/Data/Tornado/

[10] http://www.csc.kth.se/~weinkauf/notes/cylinder2d.html

[11] http://cmrm.med.jhmi.edu/cmrm/page_register/registrationORcheckpw_ok.html

[12] http://www.humanconnectomeproject.org/data/

[13] http://adni.loni.usc.edu/data-samples/

[14] http://projects.iq.harvard.edu/dtichallenge15/home

manipulate or deform measured data to generate ensembles of new datasets, e.g., by employing statistical (*random*) and physically-based (*simulation*-based) deformations [47].

Rather simple synthetic datasets are only used to demonstrate or test certain properties of developed methods. Examples include synthetic data with separating or crossing fiber structures [50, 62] and synthetic data consisting of a disk with varying diffusivity and anisotropy [61]. Synthetic datasets are used to test tensor fiber tracking [68, 113]. Simple *simulated* stress tensor fields are applied for hyperstreamline techniques [25], and for 3D tensor field topology [48]. Another synthetic tensor field is based on *randomly* generated tensors placed at the eight corners of a 3D cell, with trilinear interpolation within the cell [130].

There are also more complex approaches for generating DTI data, e.g., the generation of synthetic DTI data of a human brain with a tumor [94, 96]. In these cases, the growth of tumors is simulated including biochemical processes, i.e., the approaches use rather complex *simulation* concepts. In engineering applications, most use cases already involve simulations. Hence, respective simulation tools already exist and have not to be developed for the purpose of evaluation. Examples are simulated stress tensor fields [27, 52, 131] or gradient tensor fields derived from computational fluid dynamics (CFD) [129], or from simulations of earthquakes and engines [87].

3.4 Multi-Dimensional Data Visualization

Multi- and high-dimensional data visualization techniques depict visual mappings and transformations such as dimension reduction and ordering. They are typically designed to solve tasks such as visual analysis of clusters, correlations, and outliers among a subset of dimensions and records.

The outstanding characteristic of multi-dimensional data is the so-called curse of dimensionality. With increasing dimensionality, a global similarity definition between data records loses its discriminative behavior [10, 49]. Furthermore, irrelevant, redundant, and conflicting dimensions highly influence the understanding of a given dataset, hence also pose a challenge when generating data, e.g., for evaluation purposes.

While a few approaches compare the results of multiple datasets (e.g., [70, 106, 127]), most techniques evaluate only one or two datasets. Also, there is no established set of commonly used datasets or generation models. This fact influences a fair comparison.

In the literature, there are three types of “data sources” that are used in multi-dimensional data visualizations.

(1) It is common practice to use domain-specific datasets that are not openly accessible due to e.g., privacy restrictions. In a publication, the results are shown and described, but it is not possible to reproduce and compare the results with other techniques. A few methods exist to automatically generate data based on such *measured* world data [19, 30, 31, 81]. The model hereby tries to preserve the relationships of the original data such as the data distribution, correlations, and clusters. While these methods are useful when the original data is not available, it is rather uncommon to use them in the visualization community.

(2) From the data mining and machine learning community, there are various repositories of publicly available real-world datasets, often resulting from *measurements*. The datasets are usually labeled for specific (data mining) tasks such as classification, regression, and cluster analysis. Commonly used datasets comprise the Iris, Car, Yeast, Ecoli, and Wine datasets from the UCI Machine Learning Repository[15], the census data[16], or different variations of the so-called SwissRoll[17] (e.g., [97]) for the evaluation of dimension projection techniques. Another well-known repository is the Collections of Datasets by Weka[18][45].

(3) Many visualization techniques are applied to synthetic data. The main approach to generating multi-dimensional data is to determine a number of dimensions, add the corresponding patterns to the data (e.g., clusters, correlations, outliers, etc.), and overlay it with distractors such as noise and outliers. Besides the pattern model itself, many parameters influence the generation process: number of dimensions, instances, and patterns as well as pattern size, overlap of patterns, data distribution of patterns and dimensions, and the impact of the distractors (i.e., the amount of noise that is added).

Synthetic data is frequently generated for one approach and explicitly designed to show that existing methods are outperformed, sometimes even without providing a description of the underlying generation model. Even though, publications like [36] provide a high-level description of the generation model. Regrettably, except for some comparison papers such as [78], synthetic datasets are often not publicly available.

In the following, we describe commonly used approaches to generate multi-dimensional data. Many researchers write small scripts to generate data. In a few approaches the underlying model is well described and defined, e.g., by intersecting planes [72] or variations of the SwissRoll [97]. Other approaches [55, 56] use rules and statistics to encode relationships between data instances (e.g., older person implies higher income) and allow one to insert anomalies for different applications. The data is typically created in a black-box manner, making its scope and validity hard to grasp.

To overcome the black-box problem, recent approaches include the *user-defined* elements in the data generation process. While such semi-automatic approaches are typically time-consuming, the user knows exactly which and how patterns are distributed in the data. The approach by Albuquerque et al. [3] focuses on visual properties by representing user-defined structures as probability density functions. Afterward, the specified number of points is sampled according to these structures. In the PCDC tool [15], a user can visually add clusters of different size in subsets of the dimensions. The data instances are *randomly* distributed according to user-selected distributions within a cluster and each dimension. Another approach [120] lets the user sketch specific characteristics in parallel coordinate plots. iLAMP [93] starts from patterns in a low-dimensional space and projects them afterward in a high-dimensional space.

3.5 Network and Graph Visualization

Instances of graph data generally serve as illustrations for network visualization techniques, to evaluate graph layout algorithms, or as treatments in user studies.

Repositories such as SNAP,[19] KONECT,[20] the UFL Sparse

[15] http://archive.ics.uci.edu/ml/

[16] http://www.census.gov/data.html

[17] http://isomap.stanford.edu/datasets.html

[18] http://www.cs.waikato.ac.nz/ml/weka/datasets.html

[19] http://snap.stanford.edu/

[20] http://konect.uni-koblenz.de/

Matrix Collection,[21] the DIMACS Implementation Challenges,[22] and the Graph 500 List[23] are opportunistic collections of benchmark data. Many experiments in the graph literature have made use of the same two collections,[24] referred to as the Rome graphs and the AT&T graphs. None of these collections satisfies any criterion of representativeness or coverage.

Since network models are an important tool in network science, and social network research in particular, a plethora of models have been proposed [42, 109]. Although it is often claimed that the intention is to reproduce empirical networks, most of these models are highly idealized.

Like for other combinatorial structures, network models consist of a family of sets of graphs defining the possible outcomes, and a probability distribution on this state space. The classic examples are uniform *random* graphs, the so-called Erdős-Rény model $G(n, m)$, where the possible outcomes are all undirected graphs with exactly n vertices and m edges, and probabilities are uniform [34]. This is largely equivalent to the $G(n, p)$ model, in which each edge is included with probability p independently. (For $p = m/\binom{n}{2}$, the expected number of edges is m.) A graph sampled uniformly at random is likely to exhibit certain properties [12] that are hardly representative for networks in specific application domains. The $G(n, p)$ model is useful rather as a baseline model.

Structurally biased models usually fix certain graph invariants or other properties, or specify a distribution for them. For instance, the degree sequence may be fixed or drawn from a *global* distribution, or a partition may be given such that edge probabilities are uniform only for pairs of vertices from the same two sets. Models such as preferential attachment, where vertices are introduced one at a time and linked to other based on their current degree [6], are more conveniently formulated as processes.

Dependencies may also be introduced via vertex attributes such as preferential attachment based on a vertex attribute [11] or edge probabilities depending on similarity in some latent space defined by vertex attributes such as random coordinates [51]. When attributes are part of the output, they are usually drawn from some distribution after the graph has been created.

In the small-world model, a systematically constructed sparse graph with high *local* clustering is modified, using its symmetric difference with a uniform random graph, because this introduces shortcuts that reduce the average distance [121]. More generally, observed or constructed graphs can be subjected to random or biased modification to establish or prevent features, or to produce a *sequence* of graphs as input for dynamic visualization systems.

If they involve generators at all, user studies often utilize networks from simple random graph models such as $G(n, p)$ or its variants [125, 44]. Sometimes local variation is introduced, and the distribution is pre-specified [117, 80, 9]. Such specifications may even be the result of users sketching adjacency matrices [124]. Properties other than clustering and degrees distribution are rarely controlled despite their likely relevance in shaping a visualization.

The crucial problem with all of these generators is that they are defined with attention to a bounded number of graph features and therefore highly idealized. Samples generated from them are rarely representative and often exhibit unwanted collateral features. A few selected features do not characterize all relevant graphs, and samples need not even be realistic. Preferential attachment, for example, generates scale-free networks with characteristic power-law degree distribution, but only a tiny subset of them [91].

Two usage scenarios hold promise for better control of input factors. On the one hand, coverage of the experimental region can be controlled, for example, by rejecting samples that are too close to existing ones, and on the other hand, a small set of available network observations can be enlarged, for example, by perturbing the benchmark instances using network models that create additional instances in their vicinity and thus reduce unknown biases. As a concrete example, consider the generation of evolving social networks similar to an observed one. An exponential random graph model [74] can be fitted to the initial observation, and a stochastic actor-oriented model [110] to its evolution. Given the estimated parameters, any number of additional sequences can be generated by applying the evolution model to samples from the static model [14].

3.6 Trajectory Visualization

Trajectory visualization plays an important role in the analysis of movement [26]. It allows us to understand movement patterns of various kinds of moving "objects" in a geospatial context. Typical examples include moving vehicles, persons, or animals. The latter is investigated in movement ecology [108], with a growing amount of data available in databases.[25] The analysis of spatio-temporal eye-tracking data is yet another example, albeit with less data publicly available [63]. Whereas most types of trajectories are related to motion in 2D or 3D space, there are other trajectories that live in multidimensional state spaces, or phase spaces, where the change of values can be seen as trajectories [43].

In studies, it is important to cover the whole parameter space with the stimuli that are used, but often there is not enough *measured* real-world data that exhibit the necessary characteristics. Therefore, measurements can be used to drive a realistic *emulation* of data. For example, a Markov chain model produces synthetic trajectories from real-world movement GPS data from birds [79] that can be used to assess different visual direction encodings of trajectories.

The approach of training such an emulation model is not always possible due to insufficient data. As an alternative, methods are employed that generate purely synthetic data. For example, there are *closed-form* descriptions of semantic-based trajectories of moving objects that incorporate *random* characteristics [92, 116].

A shortcoming of many trajectory generation methods is that they only produce rather rough trajectories. Smooth trajectories can be obtained by *simulation* that adopts attraction–repulsion interaction controlled by object distances [101]. The analysis of movement data of multiple objects often involves the extraction and abstraction of common patterns, which could be visualized by using networks. According object movement data can be based on a traffic *simulation* utilizing a given network [17]. Here, objects are moving from a start location to a destination. In each time step of the *simulation*, objects are generated, moved, or removed, if they reached their destination.

[21] http://www.cise.ufl.edu/research/sparse/matrices/

[22] http://dimacs.rutgers.edu/Challenges/

[23] http://www.graph500.org/

[24] http://www.graphdrawing.org/

[25] https://www.movebank.org/, http://www.uva-bits.nl/

Eye-tracking data is conceptually similar to movement data because both types of data describe trajectories in space and time. However, there are a number of differences [4] that require different generative models: eye-movement data are recorded on a smaller spatial and temporal scale, and there are “jumps” (saccades) between fixations that are extremely fast and during which there is no cognitive processing. There is a model to *simulate* eye-tracking data [32]. The only input to this model is a list of coordinates that participants would look at during an experiment. The rest of the gaze behavior is *emulated* according to parametrized noise (*random*) functions covering the spatio-temporal fixation perturbation (modeling of a fixation), saccade velocity, and the control of the time step and sampling rate.

3.7 Text Visualization

Many different disciplines work with texts. The text analysis and the concomitant visualizations differ by discipline and according to the task at hand. Text analysis can be concerned with comparatively “surface” type properties such as the number or character of words, letters, or sentences, or can take the form of deep linguistic analysis. The former is based on opportunistic data collection that often also raises privacy/copyright issues, the models resulting from deep linguistic analysis have traditionally been deterministic generative rewriting systems and in recent years have been augmented by machine learning systems.

Linguistic analyses involve complex theoretical concepts and tend to span multiple layers of information and representation. Within computational linguistics and natural language processing (NLP), trees, and attribute-value matrices (AVM) were established early on for the representation of hierarchical and dependency relations. These tend to be stored as nested lists, but can also be visualized as directed acyclic graphs. A recent example of a sophisticated infrastructure for the storage (banking), visualization, and interactive working with complex linguistic analyses is the INESS *Infrastructure for the Exploration of Syntax and Semantics* [75, 99],[26] which is part of CLARIN,[27] the *Common Language Resources and Technology Infrastructure*, which in turn provides support and resources for research in the humanities and social sciences that is concerned with digital language data. INESS augments existing tree and AVM representations with further visualization possibilities that link and align information across complex structures and support disambiguation. The structures banked in systems like INESS range from manually constructed to fully automatically generated. The automatic generation can be the result of sophisticated manually constructed *rewrite* systems (generative grammars), various machine learning algorithms, or a combination of both. Linguistic information ranges from comparatively simple information like Part-of-Speech tags to complex semantic and pragmatic annotation, including discourse structure. In recent years, several different visualization systems for these subfields have begun to be developed [16, 23, 128].[28]

Part-of-Speech tagging along with basic morphological analysis can now be done effectively for a range of languages (again, via rule-based systems, machine learning algorithms or a combination thereof) and tend to form the input for visualizations used in other fields in the humanities and social sciences as well as for computational applications. However, the bulk of existing computational applications and increasingly, new approaches within digital humanities and social sciences work with linguistically unannotated and opportunistically collected texts. The analyses focus on surface properties such as word/sentence/text length or type/token ratios. Word clouds [59] continue to be popular, but more sophisticated visual analyses of text tend to look at several different properties in comparison and do use linguistic information [18, 60, 83, 84].

Work within the newly emergent fields of digital humanities and social sciences is also increasingly making use of generative models for the clustering and classification of texts and documents [1]. In particular, named-entity recognition [111], sentiment analysis [88], and topic modeling are being experimented with. Topic modeling in combination with visualization methods has been particularly attractive for literary analysis [54, 57, 76, 77]. One of the pioneering works for visual exploration of topic modeling results is *TIARA* [122]. This tool utilizes a theme-river visualization to show the temporal trends in topics. As it is one of the first systems that enabled the visual exploration of the results of the generative LDA model, showing the evolution of topics over time, it is considered an inspiration for many applications that have evolved afterward. Such systems include *TextFlow* [22], *Paralleltopics* [29], and *ConToVi* [33] as an application for the digital social sciences.

Beyond the visualization of text and document properties, work has recently begun to move toward providing interactive visual analytic access to complex interdependencies between linguistic and extralinguistic properties of the written material. One example is a pixel visualization of properties extracted from the annotated digital historical corpus of Icelandic [98], whereby data is generated on the basis of the annotation, but the visualization reflects a further level of abstraction from the underlying data [103, 104]. Another example is work conducted on understanding argumentation in political debates [41], where a variety of different types of information can be accessed, visualized, and explored interactively. Many of the properties that are accessed from the underlying data are the result of sophisticated linguistic and computational analysis and have been produced via generative models of various kinds. These include both rewriting systems and imitation systems.

In sum, text visualization spans a variety of different disciplines and tasks and works with various types of data, ranging from opportunistically collected and manually annotated to generatively and systematically produced. Concomitantly, a large range of visualization options have also been developed and continue to be developed, particularly in the emerging fields of digital humanities and social sciences.

4. FUTURE DIRECTIONS

A highly interesting finding of our survey is that most datasets used for the assessment of visualizations are based on measurements, i.e., there is hardly any use of generative data models. We have pointed out the few directions where there are generative models. However, these often lack controllability in the sense that they would allow us to generate collections of datasets that could be used in testing visualization techniques for such a population of inputs.

[26] http://clarino.uib.no/iness

[27] https://www.clarin.eu/content/about-clarin

[28] http://ling.uni-konstanz.de/pages/home/butt/main/material/lingvis/ for a more complete overview.

Nevertheless, we argue that there is an increasing demand for such controlled input generation if we want to make progress with empirical methods in visualization research. Therefore, we see the need for more and better use of generative data models in our community. In particular, we see considerable demand for best-practice examples and guidelines for systematic generation of problem instances across visualization domains. In the following, we highlight some specific problem areas and concrete directions for future research.

4.1 User Studies

User studies constitute the subdomain of visualization research in which experimental design is most routine and advanced. However, user studies often consider instances obtained from generative models as the treatments. Since subjects are confronted with visualization artifacts representing data, rather than the data itself, it should not be taken for granted that the variation of factors is still systematic after the samples have undergone the transformations by the visualization technique. As a further complication, interactions between the parameters of a generative model and those of a visualization technique may introduce confounding factors that are difficult to control. More research is therefore needed on what constitutes a suitable set of treatments, and how to obtain them. Given the low cost of instance generation, for instance, we see potential for the study of methods that oversample data to be able to select sets of visualization artifacts that are more suitable as a collection of treatments.

4.2 Data and Parameter Characteristics

In most examples surveyed above, a key component in the design of a generative model is the identification of data characteristics to consider, and how to make them depend on parameters. Especially in simulation, the implications of parametrization may be anything but straightforward. The parameters of a generative model are rarely independent. Ignorance of dependencies among otherwise systematically varied parameters may result in uneven or even incomplete coverage of the experimental region. Controlled studies therefore need to take into account also those interaction effects that are introduced by, and possibly specific to, the generator. Generative models are usually designed to yield data samples with special characteristics; on the flip side, this makes them prone to introduce systematic biases. Even if the subtleties of a model specification are well understood, it may be challenging to implement a generator that is correct and efficient.

4.3 Visual Mapping Characteristics

We have argued that the main purpose of generative models is to establish associations between input characteristics on the one hand and performance or visualization outcomes on the other. Standard experiments yield empirical response curves, such as running time versus problem size. Sensitivity of visualization outcomes to small perturbations in the input and other relationships may be of interest, too, especially for interactive systems. The design of generative models with parameters specifically introduced for more general forms of associations are therefore of genuine interest. We note that the utility of response curves for dependent variables such as representation accuracy or image characteristics extends beyond the assessment of visualization systems as it may be used to control the variation of factors in user studies more systematically.

4.4 Scaling

When empirical data is scarce or too small to uncover the behavior of a visualization technique, generating larger data with similar characteristics requires some form of extrapolation. Conversely, iterative development of resource-intensive visualization techniques may require smaller data, again with similar characteristics. Understanding normalization and scaling effects may thus be an important issue in the design of generative models. In large-scale simulations and other high-throughput scenarios, we might also be interested generators that emulate data from samples of a data stream.

4.5 Verification

Visualization algorithms should be subject to the same verification process that is used in other components of the scientific pipeline—this is also called *verifiable visualization* [35]. We expect that semi-automatic verification techniques, such as *fuzzing*, could greatly benefit from generative models due to extra knowledge about the parameter space.

4.6 Replication

Generative models facilitate follow-up experiments that may shed additional light on findings all too often explained by some plausible, but untested, interpretation of experimental outcomes. Who can carry out replications, follow-ups, and extensions depends on the availability of a generator. Systematic biases introduced by differences in implementation or hardware environment are just as relevant as understanding the level of documentation necessary to build an equivalent generator. When is a set of generated instances needed to reproduce results, what is the influence of random number generators (the most basic generative models), and is there a trade-off between the computational efficiency of a generation and its usability in an experiment? Archiving and versioning generative models may be a task as huge as it is for benchmark data.

5. CONCLUSION

We have made an argument for more use of generative data models in controlled visualization experiments, not just user studies, but also studies of technical performance. To support our position, we have surveyed the state of generative data models for several visualization domains. Based on our findings, we have outlined areas of interest and directions for future research.

6. ACKNOWLEDGMENTS

We thank the German Research Foundation (DFG) for financial support within Task Force 2 of SFB/Transregio 161 and project A1 of SFB-TRR 75.

7. REFERENCES

[1] C. C. Aggarwal and C. Zhai. *Mining Text Data*. Springer Science & Business Media, 2012.

[2] R. Agrawal, A. Kadadi, X. Dai, and F. Andrès. Challenges and opportunities with big data visualization. In *Proceedings of the 7th International Conference on Management of Computational and*

Collective Intelligence in Digital EcoSystems, pages 169–173, 2015.

[3] G. Albuquerque, T. Löwe, and M. Magnor. Synthetic generation of high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2317–2324, 2011.

[4] G. Andrienko, N. Andrienko, M. Burch, and D. Weiskopf. Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2889–2898, 2012.

[5] G. R. Baker. Localization: Conventional and CT simulation. *The British Journal of Radiology*, 79(1):S36–S49, 2006.

[6] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[7] W. L. Barth and C. A. Burns. Virtual rheoscopic fluids for flow visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1751–1758, 2007.

[8] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. The state of the art in visualizing dynamic graphs. In R. Borgo, R. Maciejewski, and I. Viola, editors, *EuroVis – STARs*. The Eurographics Association, 2014.

[9] M. Behrisch, B. Bach, N. H. Riche, T. Schreck, and J.-D. Fekete. Matrix reordering methods for table and network visualization. *Computer Graphics Forum*, 35(3):693–716, 2016.

[10] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is “nearest neighbor” meaningful? In *Proceedings of the 7th International Conference on Database Theory*, ICDT ’99, pages 217–235. Springer-Verlag, 1999.

[11] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436, 2001.

[12] B. Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001.

[13] R. P. Botchen, D. Weiskopf, and T. Ertl. Texture-based visualization of uncertainty in flow fields. In *IEEE Visualization Conference*, pages 647–654, 2005.

[14] U. Brandes and M. Mader. A quantitative comparison of stress-minimization approaches for offline dynamic graph drawing. In M. J. van Kreveld and B. Speckmann, editors, *Proceedings of the 19th International Symposium on Graph Drawing (GD 2011)*, volume 7034 of *Lecture Notes in Computer Science*, pages 99–110. Springer, 2011.

[15] S. Bremm, M. Heß, T. von Landesberger, and D. W. Fellner. PCDC – on the highway to data – a tool for the fast generation of large synthetic data sets. In K. Matkovic and G. Santucci, editors, *EuroVA 2012: International Workshop on Visual Analytics*. The Eurographics Association, 2012.

[16] S. Bremm, T. von Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher. Interactive visual comparison of multiple trees. In *IEEE Visual Analytics Science and Technology (VAST)*, pages 31–40, 2011.

[17] T. Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.

[18] M. Büchler, G. Heyer, and S. Gründer. eAQUA — bringing modern text mining approaches to two thousand years old ancient texts. In *Proceedings of the 4th International Conference on e-Science (IEEE08)*, 2008.

[19] G. Caiola and J. P. Reiter. Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3(1):27–42, 2010.

[20] R. Castro and E. Chuvieco. Modeling forest fire danger from geographic information systems. *Geocarto International*, 13(1):15–23, 1998.

[21] R. A. Crawfis. *New Techniques for the Scientific Visualization of Three-Dimensional Multi-Variate and Vector Fields*. PhD thesis, University of California Davis, 1995.

[22] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011.

[23] C. Culy and V. Lyding. Double tree: An advanced KWIC visualization for expert users. In *14th International Conference on Information Visualisation*, pages 98–103, 2010.

[24] B. Cutler, J. Dorsey, L. McMillan, M. Müller, and R. Jagnow. A procedural approach to authoring solid models. *ACM Transactions on Graphics*, 21(3):302–311, 2002.

[25] T. Delmarcelle and L. Hesselink. Visualizing second-order tensor fields with hyperstreamlines. *IEEE Computer Graphics and Applications*, 13(4):25–33, 1993.

[26] U. Demšar, K. Buchin, F. Cagnacci, K. Safi, B. Speckmann, N. Van de Weghe, D. Weiskopf, and R. Weibel. Analysis and visualisation of movement: an interdisciplinary review. *Movement Ecology*, 3(1):1–24, 2015.

[27] C. Dick, J. Georgii, R. Burgkart, and R. Westermann. Stress tensor field visualization for implant planning in orthopedics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1399–1406, 2009.

[28] Y. Dong, S. Lefebvre, X. Tong, and G. Drettakis. Lazy solid texture synthesis. In *Proceedings of the Nineteenth Eurographics Conference on Rendering*, EGSR ’08, pages 1165–1174. Eurographics Association, 2008.

[29] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *IEEE Visual Analytics Science and Technology (VAST)*, pages 231–240, 2011.

[30] J. Drechsler. Using support vector machines for generating synthetic datasets. In J. Domingo-Ferrer and E. Magkos, editors, *Privacy in Statistical Databases*, number 6344 in Lecture Notes in Computer Science, pages 148–161. Springer Berlin Heidelberg, 2010.

[31] J. Drechsler and J. P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243, 2011.

[32] A. T. Duchowski, S. Jörg, T. N. Allen, I. Giannopoulos, and K. Krejtz. Eye movement synthesis. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16, pages 147–154, 2016.

[33] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim. ConToVi: multi-party conversation exploration using topic-space views. *Computer Graphics Forum*, 35(3):431–440, 2016.

[34] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[35] T. Etiene, C. E. Scheidegger, L. G. Nonato, R. M. Kirby, and C. T. Silva. Verifiable visualization for isosurface extraction. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1227–1234, 2009.

[36] B. J. Ferdosi, H. Buddelmeijer, S. Trager, M. H. Wilkinson, and J. B. Roerdink. Finding and visualizing relevant subspaces for clustering high-dimensional astronomical data using connected morphological operators. In *IEEE Visual Analytics Science and Technology (VAST)*, pages 35–42, 2010.

[37] M. Üffinger, S. Frey, and T. Ertl. Interactive high-quality visualization of higher-order finite elements. *Computer Graphics Forum*, 29(2):337–346, 2010.

[38] S. Frey, G. Reina, and T. Ertl. SIMT microscheduling: Reducing thread stalling in divergent iterative algorithms. In *Proceedings of the Euromicro International Conference on Parallel, Distributed and Network-based Processing*, pages 399–406. IEEE Computer Society, 2012.

[39] S. Frey, F. Sadlo, and T. Ertl. Visualization of temporal similarity in field data. *IEEE Transactions on Visualization and Computer Graphics*, 18:2023–2032, 2012.

[40] C. Garth and X. Tricoche. Topology- and feature-based flow visualization: Methods and applications. In *SIAM Conference on Geometric Design and Computing*, 2005.

[41] V. Gold, M. El-Assady, T. Bögel, C. Rohrdantz, M. Butt, K. Holzinger, and D. Keim. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 2015.

[42] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2009.

[43] S. Grottel, J. Heinrich, D. Weiskopf, and S. Gumhold. Visual analysis of trajectories in multi-dimensional state spaces. *Computer Graphics Forum*, 33(6):310–321, 2014.

[44] H. Guo, J. Huang, and D. H. Laidlaw. Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1173–1186, 2015.

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[46] G. Haller. Lagrangian coherent structures. *Annual Review of Fluid Mechanics*, 47:137–162, 2015.

[47] G. Hamarneh, P. Jassi, and L. Tang. *Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008, Part I*, chapter Simulation of Ground-Truth Validation Data Via Physically- and Statistically-Based Warps, pages 459–467. Springer Berlin Heidelberg, 2008.

[48] L. Hesselink, Y. Levy, and Y. Lavin. The topology of symmetric, second-order 3D tensor fields. *IEEE Transactions on Visualization and Computer Graphics*, 3(1):1–11, 1997.

[49] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 506–515. Morgan Kaufmann Publishers Inc., 2000.

[50] M. Hlawatsch, J. E. Vollrath, F. Sadlo, and D. Weiskopf. Coherent structures of characteristic curves in symmetric second order tensor fields. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):781–794, 2011.

[51] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[52] I. Hotz, L. Feng, H. Hagen, B. Hamann, K. Joy, and B. Jeremic. Physically based methods for tensor field visualization. In *IEEE Visualization Conference*, pages 123–130, 2004.

[53] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[54] S. Jänicke, G. Franzini, M. Cheema, and G. Scheuermann. On close and distant reading in digital humanities: A survey and future challenges. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) – STARs*. The Eurographics Association, 2015.

[55] D. R. Jeske, P. J. Lin, C. Rendon, R. Xiao, and B. Samadi. Synthetic data generation capabilties for testing data mining tools. In *IEEE Military Communications Conference MILCOM 2006*, pages 1–6, 2006.

[56] D. R. Jeske, B. Samadi, P. J. Lin, L. Ye, S. Cox, R. Xiao, T. Younglove, M. Ly, D. Holt, and R. Rich. Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 756–762. ACM, 2005.

[57] M. L. Jockers. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press, 2013.

[58] G. K. Karch, F. Sadlo, D. Weiskopf, C.-D. Munz, and T. Ertl. Visualization of advection-diffusion in unsteady fluid flow. *Computer Graphics Forum*, 31(3):1105–1114, 2012.

[59] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization. In *Proceedings of*

the Workshop on Tagging and Metadata for Social Information Organization at WWW2007, 2007.

[60] D. A. Keim and D. Oelke. Literature fingerprinting: A new method for visual literary analysis. In *IEEE Visual Analytics Science and Technology (VAST)*, pages 115–122, 2007.

[61] G. Kindlmann, D. B. Ennis, R. T. Whitaker, and C. F. Westin. Diffusion tensor analysis with invariant gradients and rotation tangents. *IEEE Transactions on Medical Imaging*, 26(11):1483–1499, 2007.

[62] G. Kindlmann, X. Tricoche, and C.-F. Westin. Delineating white matter structure in diffusion tensor MRI with anisotropy creases. *Medical Image Analysis*, 11(5):492–502, 2007.

[63] K. Kurzhals, C. F. Bopp, J. Bässler, F. Ebinger, and D. Weiskopf. Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, BELIV '14, pages 54–60. ACM, 2014.

[64] D. H. Laidlaw, R. M. Kirby, C. D. Jackson, J. S. Davidson, T. S. Miller, M. da Silva, W. H. Warren, and M. J. Tarr. Comparing 2D vector field visualization methods: A user study. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):59–70, 2005.

[65] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[66] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf. The state of the art in flow visualization: Dense and texture-based techniques. *Computer Graphics Forum*, 23(2):203–221, 2004.

[67] R. S. Laramee, H. Hauser, L. Zhao, and F. H. Post. Topology-based flow visualization, the state of the art. In *Topology-based Methods in Visualization*, pages 1–19. Springer Berlin Heidelberg, 2007.

[68] M. Lazar and A. L. Alexander. An error analysis of white matter tractography methods: Synthetic diffusion tensor field simulations. *Neuroimage*, 20(2):1140–1153, 2003.

[69] D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546, 2001.

[70] D. J. Lehmann and H. Theisel. Optimal sets of projections of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):609–618, 2016.

[71] G.-S. Li, X. Tricoche, and C. Hansen. Physically-based dye advection for flow visualization. *Computer Graphics Forum*, 27(3):727–734, 2008.

[72] S. Liu, P. T. Bremer, J. J. Jayaraman, B. Wang, B. Summa, and V. Pascucci. The Grassmannian atlas: A general framework for exploring linear projections of high-dimensional data. *Computer Graphics Forum*, 35(3):1–10, 2016.

[73] Z. Liu, S. Cai, J. E. Swan, R. J. Moorhead, J. P. Martin, and T. J. Jankun-Kelly. A 2D flow visualization user study using explicit flow synthesis and implicit task design. *IEEE Transactions on Visualization and Computer Graphics*, 18(5):783–796, 2012.

[74] D. Lusher, J. Koskinen, and G. Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications.* Cambridge University Press, 2012.

[75] P. Meurer, V. Rosén, and K. D. Smedt. Interactive visualizations in the INESS treebanking infrastructure. In A. Hautli-Janisz and V. Lyding, editors, *Proceedings of the LREC 2016 Workshop "VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources"*, pages 1–7, 2016.

[76] F. Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, 2005.

[77] F. Moretti. *Distant Reading.* Verso, 2013.

[78] E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *Proceedings of the VLDB Endowment*, 2(1):1270–1281, 2009.

[79] R. Netzel, M. Burch, and D. Weiskopf. Comparative eye tracking study on node-link visualizations of trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2221–2230, 2014.

[80] A. Nocaj, M. Ortmann, and U. Brandes. Untangling the hairballs of multi-centered, small-world online social media networks. *Journal of Graph Algorithms and Applications*, 19(2):595–618, 2015.

[81] B. Nowok, G. M. Raab, and C. Dibben. *synthpop: Bespoke creation of synthetic data in R.* University of Edinburgh, 2015.

[82] J. O'Callaghan, J. Wells, S. Richardson, H. Holmes, Y. Yu, S. Walker-Samuel, B. Siow, and M. F. Lythgoe. Is your system calibrated? MRI gradient system calibration for pre-clinical, high-resolution imaging. *PLoS One*, 9(5):1–9, 2014.

[83] D. Oelke, D. Kokkinakis, and M. Malm. Advanced visual analytics methods for literature analysis. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012) at EACL 2012*, pages 35–44. Association for Computational Linguistics, 2012.

[84] D. Oelke, D. Spretke, A. Stoffel, and D. Keim. Visual readability analysis: How to make your writings easier to read. *IEEE Transactions on Visualization and Computer Graphics*, 18(5):662–674, 2012.

[85] M. Otto, T. Germer, H.-C. Hege, and H. Theisel. Uncertain 2D vector field topology. *Computer Graphics Forum*, 29(2):347–356, 2010.

[86] S. Owada, T. Harada, P. Holzer, and T. Igarashi. Volume painter: Geometry-guided volume modeling by sketching on the cross-section. In *Proceedings of the Fifth Eurographics Conference on Sketch-Based Interfaces and Modeling*, SBM'08, pages 9–16. Eurographics Association, 2008.

[87] D. Palke, Z. Lin, G. Chen, H. Yeh, P. Vincent, R. Laramee, and E. Zhang. Asymmetric tensor field visualization for surfaces. *IEEE Transactions on*

Visualization and Computer Graphics, 17(12):1979–1988, 2011.

[88] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[89] R. Peikert and F. Sadlo. *Topology-Based Methods in Visualization II*, chapter Flow Topology Beyond Skeletons: Visualization of Features in Recirculating Flow, pages 145–160. Springer Berlin Heidelberg, 2009.

[90] K. Perlin. An image synthesizer. *SIGGRAPH Computer Graphics*, 19(3):287–296, 1985.

[91] C. Petersen, N. Rotbart, J. G. Simonsen, and C. Wulff-Nilsen. Near-optimal adjacency labeling scheme for power-law graphs. *CoRR*, abs/1502.03971, 2015.

[92] D. Pfoser and Y. Theodoridis. Generating semantics-based trajectories of moving objects. *Computers, Environment and Urban Systems*, 27(3):243–263, 2003.

[93] E. Portes dos Santos Amorim, E. V. Brazil, J. Daniels, P. Joia, L. G. Nonato, and M. C. Sousa. iLAMP: Exploring high-dimensional spacing through backward multidimensional projection. In *IEEE Visual Analytics Science and Technology (VAST)*, pages 53–62, 2012.

[94] M. Prastawa, E. Bullitt, and G. Gerig. Simulation of brain tumors in MR images for evaluation of segmentation efficacy. *Medical Image Analysis*, 13(2):297–311, 2009.

[95] S. Pujol, W. Wells, C. Pierpaoli, C. Brun, J. Gee, G. Cheng, B. Vemuri, O. Commowick, S. Prima, A. Stamm, M. Goubran, A. Khan, T. Peters, P. Neher, K. H. Maier-Hein, Y. Shi, A. Tristan-Vega, G. Veni, R. Whitaker, M. Styner, C.-F. Westin, S. Gouttard, I. Norton, L. Chauvin, H. Mamata, G. Gerig, A. Nabavi, A. Golby, and R. Kikinis. The DTI challenge: Toward standardized evaluation of diffusion tensor imaging tractography for neurosurgery. *Journal of Neuroimaging*, 25(6):875–882, 2015.

[96] J. Rexilius, H. K. Hahn, M. Schlüter, S. Kohle, H. Bourquain, J. Böttcher, and H.-O. Peitgen. *Proceedings of the 7th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004, Part II*, chapter A Framework for the Generation of Realistic Brain Tumor Phantoms and Applications, pages 243–250. Springer Berlin Heidelberg, 2004.

[97] B. Rieck and H. Leitte. Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum*, 34(3):431–440, 2015.

[98] E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of LREC 2012*, pages 1978–1984, 2012.

[99] V. Rosén, K. D. Smedt, P. Meurer, and H. Dyvik. An open infrastructure for advanced treebanking. In J. Hajič, K. D. Smedt, M. Tadić, and A. Branco, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, 2012.

[100] P. G. Saffman. *Vortex Dynamics*. Cambridge University Press, 1993. Cambridge Books Online.

[101] J.-M. Saglio and J. Moreira. Oporto: A realistic scenario generator for moving objects. *GeoInformatica*, 5(1):71–93, 2001.

[102] T. J. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[103] C. Schätzle and D. Sacha. Visualizing language change: Dative subjects in Icelandic. In A. Hautli-Janisz and V. Lyding, editors, *Proceedings of the LREC 2016 Workshop "VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources"*, pages 8–15, 2016.

[104] C. Schätzle, D. Sacha, and M. Butt. Diachronic visualization of dative subjects in Icelandic. Poster presentation at the Workshop on Big Data Visual Computing, *44th Annual Meeting of the Gesellschaft für Informatik*, 2014.

[105] T. Schiwietz, T. Chang, P. Speier, and R. Westermann. MR image reconstruction using the GPU. *SPIE Medical Imaging*, 2006.

[106] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3):1335–1344, 2012.

[107] S. C. Shadden, F. Lekien, and J. E. Marsden. Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. *Physica D*, 212(3-4):271–304, 2005.

[108] J. Shamoun-Baranes, E. E. van Loon, R. S. Purves, B. Speckmann, D. Weiskopf, and C. Camphuysen. Analysis and visualization of animal movement. *Biology Letters*, 8:6–9, 2012.

[109] T. A. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37:131–153, 2011.

[110] T. A. Snijders, G. G. van de Bunt, and C. E. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 2010.

[111] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[112] M. Tory and T. Möller. Rethinking visualization: A high-level taxonomy. In *10th IEEE Symposium on Information Visualization (InfoVis 2004)*, pages 151–158, 2004.

[113] J.-D. Tournier, F. Calamante, M. King, D. Gadian, and A. Connelly. Limitations and requirements of diffusion tensor fiber tracking: An assessment using simulations. *Magnetic Resonance in Medicine*, 47(4):701–708, 2002.

[114] J.-D. Tournier, C.-H. Yeh, F. Calamante, K.-H. Cho, A. Connelly, and C.-P. Lin. Resolving crossing fibres using constrained spherical deconvolution: Validation using diffusion-weighted imaging phantom data. *Neuroimage*, 42(2):617–625, 2008.

[115] H. Turbell. *Cone-Beam Reconstruction Using Filtered Backprojection*. PhD thesis, Linköping University, 2001.

[116] T. Tzouramanis, M. Vassilakopoulos, and Y. Manolopoulos. On the generation of time-evolving

regional data. *GeoInformatica*, 6(3):207–231, 2002.

[117] F. van Ham and M. Wattenberg. Centrality based visualization of small world graphs. *Computer Graphics Forum*, 27(3):975–982, 2008.

[118] J. J. van Wijk. Image based flow visualization. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pages 745–754. ACM, 2002.

[119] W. von Funck, T. Weinkauf, H. Theisel, and H.-P. Seidel. Smoke surfaces: An interactive flow visualization technique inspired by real-world flow experiments. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1396–1403, 2008.

[120] B. Wang, P. Ruchikachorn, and K. Mueller. SketchPadN-D: WYDIWYG sculpting and editing in high-dimensional space. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2060–2069, 2013.

[121] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[122] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. TIARA: A visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 153–162. ACM, 2010.

[123] D. Weiskopf. *GPU-Based Interactive Visualization Techniques*. Springer Berlin Heidelberg, 2006.

[124] P. C. Wong, H. Foote, P. Mackey, K. Perrine, and G. Chin. Generating graphs for visual analytics through interactive sketching. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1386–1398, 2006.

[125] K. Xu, C. Rooney, P. Passmore, D. H. Ham, and P. H. Nguyen. A user study on curved edges in graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2449–2456, 2012.

[126] N. Yanasak and J. Allison. Use of capillaries in the construction of an MRI phantom for the assessment of diffusion tensor imaging: Demonstration of performance. *Magnetic Resonance Imaging*, 24(10):1349–1361, 2006.

[127] X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu. Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1001–1008, 2009.

[128] J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan. Facilitating discourse analysis with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2639–2648, 2012.

[129] X. Zheng and A. Pang. HyperLIC. In *IEEE Visualization Conference*, pages 249–256, 2003.

[130] X. Zheng and A. Pang. Topological lines in 3D tensor fields. In *IEEE Visualization Conference*, pages 313–320, 2004.

[131] V. Zobel, M. Stommel, and G. Scheuermann. Feature-based tensor field visualization for fiber reinforced polymers. In *2015 IEEE Scientific Visualization Conference (SciVis)*, pages 49–56, 2015.