Who leads? Who follows? Temporal dynamics of political dogwhistles in Swedish online communities

Max Boholm,[1] **Gregor Rettenegger,**[2] **Ellen Breitholtz,**[3]
Robin Cooper,[3] **Elina Lindgren,**[4] **Björn Rönnerstrand,**[2] and **Asad Sayeed**[3]
[1]School of Public Administration, [2]Journalism Media and Communication (JMG),
[3]Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg
[4]Dept. of Political, Historical, Religious and Cultural Studies, Karlstad University
{max.boholm, asad.sayeed}@gu.se

Abstract

A dogwhistle is a communicative act intended to broadcast a message only understood by a select in-group while going unnoticed by others (out-group). We illustrate that political dogwhistle behavior in a more radical community precedes the occurrence of the dogwhistles in a less radical community, but the reverse does not hold. We study two Swedish online communities – *Flashback* and *Familjeliv* – which both contain discussions of life and society, with the former having a stronger anti-immigrant subtext. Expressions associated with dogwhistles are substantially more frequent in Flashback than in Familjeliv. We analyze the time series of changes in *intensity* of three dogwhistle expressions (DWEs), i.e., the strength of association of a DWE and its in-group meaning modeled by Swedish Sentence-BERT, and model the dynamic temporal relationship of intensity in the two communities for the three DWEs using Vector Autoregression (VAR). We show that changes in intensity in Familjeliv are explained by the changes of intensity observed at previous lags in Flashback but not the other way around. This suggests a direction of travel for dogwhistles associated with radical ideologies to less radical contexts.

1 Introduction

Political dogwhistles are "speech acts that explicitly convey a certain content to an audience, while simultaneously sending a different, concealed message to a specific subset of that audience" (Lo Guercio and Caso, 2022). Since dogwhistles enable communication of controversial views to sympathizers while not alienating a wider audience, dogwhistles are an efficient strategy in political communication to mobilize support, such as votes (White, 2007; Hurwitz and Peffley, 2005; Wetts and Willer, 2019; Lindgren et al., 2024; Albertson, 2015). Dogwhistles thus pose a problem for democracy by obscuring electoral mandates (Goodin and Saward, 2005; Howdle, 2023).

Additionally, dogwhistles are used online by citizens (Åkerlund, 2022; Bhat and Klein, 2020). Since dogwhistles are used for concealed expressions of intolerant discourse (Rossini, 2020) while evading accountability, they align with "dark" participation online (Lutz and Hoffmann, 2017; Quandt, 2018), such as hate speech and disinformation (Lorenz-Spreen et al., 2022). There is a growing interest in applying computational methods to such behaviors (e.g. Mendelsohn et al., 2023; Hertzberg et al., 2022; Ribeiro et al., 2020).

Since the words used in a dogwhistle have a "conventional" out-group meaning, vector-space measures of meaning change can be used to determine when it is more likely that a dogwhistle is being used with an in-group meaning. Recognizing the temporal nature of dogwhistles (Sayeed et al., 2024), recent studies have modeled the meaning change in dogwhistle expressions (DWEs) over time. They showed that the rate of change of DWEs diverge between communities (Boholm and Sayeed, 2023) and that the variability of dogwhistle *intensity*, i.e., the strength of association between DWEs and their in-group meaning, is predictable from the general patterns of semantic change of these expressions (Boholm et al., 2024).

Previous studies of dogwhistle behavior as a form of semantic change use yearly time series data, which is of limited length. With finer granularity, the non-deterministic nature of dogwhistle behavior becomes apparent. That is, while DWEs online often show a trending behavior (both in terms of frequency and intensity), there is a fluctuating pattern where values rise and fall cyclically relative to a baseline (see Figure 1). Such a pattern can be modeled as a function of the external forces that cause the series to vary with other temporal features, such as auto-regression (Box-Steffensmeier et al., 2014). Little is currently known of what explains such patterns of dogwhistle behavior online.

Our aim is to explore one factor in explaining

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 383–395

variation of dogwhistle behavior online, namely how past changes in one community could explain present changes in dogwhistle behavior in another community. We study the communities of two discussion forums: *Flashback* and *Familjeliv* (Family life). The main content of these is general discussion about life and society, but Flashback, unlike Familjeliv, has a strong anti-immigrant subtext (Åkerlund, 2021; Blomberg and Stier, 2019; Malmqvist, 2015). Thus, Flashback is expected to have a larger proportion of anti-immigrant in-group members than Familjeliv. Unsurprisingly, DWEs are more frequent in Flashback than in Familjeliv, see Table 1. Flashback and Familjeliv are widely known in Swedish society as drivers of public discourse, particularly Flashback. Flashback has served as a gateway to radicalization when dogwhistles that emerged in far-right online communities were eventually adopted in mainstream media and public discourse by politicians (Åkerlund, 2022).

Accordingly, we test the following hypotheses:

> **(H1)** (past) changes in the intensity of dogwhistles in Flashback predict changes in the intensity of dogwhistles in Familjeliv

but

> **(H2)** (past) changes in the intensity of dogwhistles in Familjeliv do *not* predict changes in the intensity of dogwhistles in Flashback

We find partial support for these. Our findings confirm previous work on the mainstreaming of hate speech online that shows that radicalization of far-right ideas is mediated by semi-radical settings such as Flashback (Klein, 2012; Ribeiro et al., 2020; Åkerlund, 2022, 2020).

2 Related work

The use of dogwhistles online can be seen as a sub-process of the radicalization of mainstream media (Åkerlund, 2022). The features of the Internet have provided subversive social movements with tools to normalize hate speech (Munn, 2019). By mimicking legitimate sources of information like news media, extremist movements have succeeded in legitimizing their causes (Klein, 2012). In this mainstreaming of hate, the role of "gateways" has been explored (Åkerlund, 2022; Mamié et al., 2021). Radicalization evolves through "pipelines" (Munn, 2019) of interactions in increasingly extreme communities (Ribeiro et al., 2020).

In her detailed, two-decade case study of the dogwhistle "cultural enricher" across the Swedish-speaking Internet, Åkerlund (2022) showed that Flashback served as a gateway to mainstream radicalization, by mediating far-right discourse. Framing content in neutral ways (Åkerlund, 2020) and concealing racist sentiment are key strategies to inject extremist ideas into less radical contexts.

Although still limited, there is growing interest in computational methods for studying dogwhistles, including formal semantics (Breitholtz and Cooper, 2021; Henderson and McCready, 2018, 2024) and data annotation (Xu et al., 2021; Kruk et al., 2024). Using data from a replacement test in which subjects were instructed to replace DWEs in sentence contexts with what they *thought* they meant, Hertzberg et al. (2022) showed that the Support Vector Machine (SVM) classifier could reliably separate "in-group" and "out-group" interpretations of the terms, based on their sentence embeddings (Reimers and Gurevych, 2019).

Mendelsohn et al. (2023) tested the ability of the Large Language Model (LLM) GPT-3 (Brown et al., 2020) to identify in-group meanings of dogwhistles, under various conditions. Similarly, Kruk et al. (2024) tested the ability of several contemporary LLMs to identify and define dogwhistles, showing high accuracy for all the models.

By applying computational methods of lexical semantic change (LSC) detection (Tahmasebi et al., 2021; Tang, 2018), Boholm and Sayeed (2023) found that the rate of change of DWEs in the two different online discussion diverged, thus showing that dogwhistle evolution is community dependent (cf. Quaranto, 2022; Clark, 1996). Boholm et al. (2024) showed that general measures of LSC (Tang, 2018; Tahmasebi et al., 2021) predicted the semantic change of the in-group relative to the out-group meaning.

Methodologically related to our study is the work that uses time series models to explain patterns of agenda-setting, i.e., the process of change in political priorities (Baumgartner and Jones, 1993) and media (McCombs and Shaw, 1972). Using social media data, Barberá et al. (2019) explored who is more likely to lead or follow in discussions on public issues via a vector autoregressive (VAR) model; they found that politicians tend to follow the priorities of the public.

3 Data

We study the temporal dynamics of dogwhistle meaning using two data sources: (1) a survey of Swedish citizens asked to perform a word replacement task for DWE-containing sentences (section 3.1) and (2) corpus data of the posts of the online discussion forums Flashback and Familjeliv (section 3.2).

3.1 Replacement survey

Lindgren et al. (2024) have provided data from a word replacement task that quantified variability in how individuals understand the meaning of dogwhistles. Swedish citizens (n=1780, pre-stratified in terms of age, gender, and education) were instructed to read sentences and replace a potential DWE in each sentence with one or more words so that the meaning of the sentence remained intact. Potential dogwhistle words were collected from Swedish media, of which twelve were identified and included in the survey.

The replacement task was completed by 900 panelists (participation rate of 51%). The survey responses were manually coded to determine whether the replacement words instantiated 1) the implicit dogwhistle meaning, 2) the explicit literal meaning, or 3) word(s) that could not be coded as 1 or 2.

We use this data to validate a set of three Swedish DWEs and to obtain language data from which we can model the in-group meaning of the DWEs (see next section; also, see Limitations). We chose three DWEs from the Lindgren et al. dataset that had high inter-annotator agreement (Krippendorff's $\alpha > 0.6$):

re-migration (*återvandring*), which has in-group and out-group meanings based on the (in)voluntariness of emigration to "home" countries, with a voluntary act as the out-group meaning, and 'deportation' as the in-group meaning.

enrich (*berika*), is the result of malevolent irony, in response to positive opinions on multiculturalism, where the in-group meaning is the opposite of enrichment, namely criminal and destructive activities (by immigrants).

globalist (*globalist*), which is used with several different in-group meanings, including an anti-Semitic reference to Jews, a nationalistic reference to anti-nationalists (i.e., opponents of nationalism), and a populist reference to elitism.

3.2 Corpora

We use corpus data to build diachronic embeddings of DWEs in different communities and to model the relation between these embeddings and the embeddings of DWE in-group replacements.

The discussion forum *Flashback* covers a wide range of topics that are organized in threads under 15 general sections (e.g., computers, lifestyle, and politics). As of March 13, 2024, the website claimed to have over 1.5 million users and more than 80 million posts. Hate speech is not allowed, but the website supports user anonymity, and Flashback is known for its discussion of controversial topics while allowing discriminatory or racist discourse (Åkerlund, 2021; Blomberg and Stier, 2019; Malmqvist, 2015; Cohen et al., 2022).

Familjeliv is similar to Flashback in being organized in threads of 20 general categories. It is less focused on contentious political issues and more focused on social issues and practical matters of everyday life. Familjeliv lacks Flashback's explicit support for members to say anything on controversial topics, instead focusing on family and parenting (Hanell and Salö, 2017) while including general-interest and political topics. Data for the two communities were collected from the Swedish National Language Bank[1].

We use quarterly data ranging from 2010 to 2024. The terms are much more common in Flashback than in Familjeliv (Table 1). Considering the mean normalized frequency, *globalist* is almost twenty times more common in Flashback than Familjeliv.

4 Semantic modeling

We consider a DWE to be more "intense" at a given moment if its observed context of use is more similar to the non-dogwhistle terms that represent the same in-group meaning (e.g., "re-migration" vs. "deportation"). We define the *intensity* of a DWE w at time period t as the similarity of the contextual embedding of w to the embeddings of the in-group replacements from the word replacement task (I^w; section 3.1).[2] The contextual embedding is taken from the contexts of w found during t in the corpus. I^w is converted to a single vector representation by taking the mean of the replacement task responses.

[1]Flashback data: https://spraakbanken.gu.se/en/resources/flashback-politik
Familjeliv data: https://spraakbanken.gu.se/en/resources/familjeliv-allmanna-samhalle

[2]Code for running experiments can be found at https://github.com/mboholm/dogwhistle-var-woah.

	Flashback			*Familjeliv*		
DWE	Total	*M*	*SD*	Total	*M*	*SD*
berika	20023	29.08	8.25	1134	8.06	5.35
globalist	31672	48.39	35.76	122	2.51	4.27
återvandring	12339	18.08	23.56	289	3.72	7.28

Table 1: Total frequency and mean frequency per million, per quarter

We use sentence embeddings from Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). SBERT is BERT (Devlin et al., 2019) fine-tuned for predicting the semantic similarity of two sentences. SBERT has a bi-encoder architecture to reduce the computational cost of sentence pair-regression in original BERT. Reimers and Gurevych (2019) show that a bi-encoder with fine-tuning reaches state-of-the-art performance on sentence similarity. Swedish SBERT (Rekathati, 2021) is trained with transfer learning following the procedure by Reimers and Gurevych (2020), where the objective is to make a student model (Malmsten et al., 2020)[3] (of an under-resourced language, here: Swedish) match the sentence embeddings of a high-performing teacher model[4] (developed for a well-resourced language, here: English) in a parallel corpus. The mean vector of the sentence embeddings for *w* at *t* constitutes $\overrightarrow{w}_t$.

The similarity of $\overrightarrow{w}_t$ and $\overrightarrow{I^w}$ is measured by the angular similarity (Kim et al., 2014; Noble et al., 2021). Angular distance is the normalized angle between two vectors, which can be calculated from the cosine similarity of the two vectors, see equation 1. Angular similarity is the complement of the angular distance. Accordingly, *intensity* of a DWE *w* at *t* is defined as follows:

$$intensity_t^w = 1 - \frac{arccos(sim(\overrightarrow{I^w}, \overrightarrow{w}_t))}{\pi} \quad (1)$$

Unlike cosine similarity (*sim*), which ranges from -1 to 1, angular similarity ranges from 0 to 1, where values close to 1 indicate strong similarity and values close to 0 indicate strong dissimilarity.

See Appendix A.1 for examples with high and low intensity calculated for individual cases.

[3] https://huggingface.co/KB/bert-base-swedish-cased

[4] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

5 Time series modeling

The dynamic relationship between the intensity of a dogwhistle in the two communities is modeled by vector autoregression (VAR). VAR is multivariate generalization of the univariate autoregressive (AR) model for time series analysis. Like AR models, VAR models estimate how the present value of an included variable, X_t, depending on previous values of the variable, X_{t-1}, ..., X_{t-p}, where *p* is the order of the model, i.e., the lag. Moreover, for each variable included, VAR models capture its relationship to the lagged values of other variables in the model.

By modeling the relationship between the *k* variables in the model, a VAR model can be described as a system of equations (Box-Steffensmeier et al., 2014), where the equation for each variable includes the variable's own lagged values, the lagged values of the other variables, and an error term. Formally this can be defined as:

$$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + e_t \quad (2)$$

where y_t is the collection of *k* variables, i.e., a vector or a $k \times 1$ matrix (hence the name of the method), *t* is the time ($t = 1, 2, \ldots, T$, where *T* is the length of the time series), *p* the order of the model (i.e. the number of lags modeled), *c* is a *k* length vector of intercept terms, and e_t a *k* length vector of error terms. A_i is a $k \times k$ matrix of (fixed) coefficients at lag *i* (Lütkepohl, 2005). The coefficients of the VAR models can be estimated by maximum likelihood estimator (MLE) or ordinary least squares (OLS). We use the latter, implemented by the Python module `statsmodels`. In VAR modeling there is no assumption of the direction of influence between variables (Sims, 1980). This feature is desired, since there is no previous theory or body of empirical work to guide any detailed expectations of dogwhistle behavior online (Box-Steffensmeier et al., 2014).

(a) Flashback

(b) Familjeliv

Figure 1: Intensity (dashed blue) and change of intensity (orange) in (a) Flashback and (b) Familjeliv (normalized by z-scores), with mean imputation of zero-frequency quarters (red dot).

5.1 Time series: quarterly changes in intensity

In the present study, we measure intensity over time in Flashback and Familjeliv. However, we do not model the level of intensity directly, but the change in intensity, i.e. we difference the time series before modeling (Figure 1).

DWEs are much less common in Familjeliv than Flashback (table 1). This raises a problem for the time series of *globalist* and *återvandring* in Familjeliv, because there are quarters where these DWEs are not observed. However, *berika* is observed in Familjeliv in all but one quarter from 2010 to 2024.

To model the temporal relationships between communities, we need an estimate of intensity in these zero-frequency quarters. Since we lack values for quarters where the DWEs have zero frequency, we impute missing values in the original series using two methods. The first method takes the mean value of the existing intensity values in the series to represent the intensity in zero-frequency quarters. Like any imputation, mean imputation can be problematic. In particular, it underestimates the standard error. To test our analysis with a fundamentally different type of imputation, we also explore Last Observation Carried Forward (LOCF) imputation (Niako et al., 2024). For a zero-frequency quarter, the intensity is identified as the intensity of the previous quarter with non-zero frequency. In the case that the intensity of the first quarter of the series is missing, we represent the first value by the first non-missing value of the series.

In VAR models, the variables (time series) are assumed to be stationary, i.e., the statistical properties of a variable (e.g., mean and variance) do not change with time. In a stationary time series, external forces – or "shocks" – to the system eventually lose their influence on present values, so that the system returns to a baseline (equilibrium or line of a trend). In contrast, in non-stationary data, shocks integrate into the system, building up over time and not returning to an equilibrium (Box-Steffensmeier et al., 2014). After imputation, we establish stationarity by removing the trend and integrative process of the data by differencing (Box-Steffensmeier et al., 2014). That is, for a time series Y_t, $differencing(Y_t) = Y_t - Y_{t-1}$. Moreover, variables in VAR are assumed to have the same order of integration, i.e. differencing one variable requires differencing of the other(s). Thus, we model how a *change* of intensity in a community predicts a *change* of intensity in another.

A combination of Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests is used to test for the stationary of variables (Box-Steffensmeier et al., 2014). Both of these test for the presence of a unit root in the time series (i.e., shocks have permanent impact on the mean or linear trend, so that the series does not return to an equilibrium). A process with a unit root is not stationary.

For *berika* and *globalist*, but not for *återvandring*, the imputed time series are not stationary.

However, for comparability, so that we test how *change* of intensity is temporally related in the two communities throughout the analysis, we difference the time series for all three terms (despite the *återvandring* series being stationary without differencing). Since there is a risk of over-differencing, we make sure that the differenced time series for all three terms still are stationary according to ADF and KPSS.

5.2 Model specification

We use a combination of three information-theoretic measures of model fit to determine which lag from 0 to 4 to use when estimating the VAR models (Box-Steffensmeier et al., 2014): Akaike information criterion (AIC), Schwarz's Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC). To counteract overfitting, all three measures penalize model complexity in the estimation of model fit (BIC more than HQIC and HQIC more than AIC). In cases where these measures disagree on the best-fitted model, we decide on the lag that the majority prefer. In one case, i.e. *återvandring*, with LOCF imputation, all three criteria have different preferences. In this case, we follow the preferred model according to AIC. However, in this case, the ordering of the VAR suggested by AIC ($p = 4$) and HQIC ($p = 2$) does not matter much for the general pattern of IRFs and the significance of Granger causality.

5.3 Granger causality

The results of VAR models are usually not communicated with the estimated coefficients of the model directly (Box-Steffensmeier et al., 2014). Since a VAR model has $k \times k \times p$ estimated coefficients, a reasonable overview is quickly lost with growing values of k and p. Instead, two other tools are common for communicating VAR results: Granger causality and Impulse Response Functions (IRPs). "Causality" is used here as a standard technical term of art.

A variable *X* is said to "Granger cause" another variable *Y*, if the previous values of *X* provide significant information in forecasting *Y*. More precisely, a test of Granger causality tests whether the past values of *Y* together with the past values of *X* enable significantly better predictions of *Y* than the past values of *Y* alone (Box-Steffensmeier et al., 2014). The procedure for testing whether *X* Granger causes *Y* is two compare the error terms of a *restricted* VAR model which estimates *Y* excluding lagged values of *X* with an *unrestricted* model which includes lagged values of *X*. If the unrestricted model is significantly better (has lower residuals) than the restricted model, it can be concluded that the past values of *X* enable significantly better predictions of *Y* than the past values of *Y* alone. The model difference (improvement) can be estimated with the F-statistic.

5.4 Impulse Response Functions (IRFs)

An IRF is a way to model how shocks of variables propagate throughout the VAR system. IRFs can be interpreted as the effect of a shock in *X* on Y (if any) and are visualized as line plots, indicating both direction and how sustained the effects of shocks are in the system. The basis of IRFs is that any VAR(p) model can be expressed as a VAR(1) model, the moving average representation of a VAR model. IRFs are estimated from "impact multipliers" derived from simplification of the moving average version of a model (Box-Steffensmeier et al., 2014, 113–115).

6 Results

In total we estimate six models: two types of imputation for the three terms (but note that for *berika* the two versions are basically the same since there are only two missing values). We here present the results for the mean imputed data in more detail, but refer to Appendix A.2 for details on the LOCF-imputed data. In general, the results of the two methods implies the same conclusions, but there are some differences related to *globalist*, which are mentioned below.

Given the complexity of many VAR models, the model coefficients are seldom communicated in studies. However, in our case, an overview of the model is possible (Table 2). Before turning to the more specific results of the IRFs and Granger causality, there are three relevant observations of the model coefficients (table 2). First, for all three DWEs, the lagged changes in intensity in Flashback are significant predictors of changes in intensity in Familjeliv. However, while the coefficients for these associations in the case of *berika* and *återvandring* are positive, the significant coefficient of the variable is negative in the case of *globalist*. Moreover, the pattern observed here for the mean-imputed data in the case of *berika* and *återvandring* is also observed for the LOCF data, but for *globalist* the coefficients of variables FB_{t-1} and

FB_{t-2} of equation FL_t are not significant (Table A1). Together, these observations partially support hypothesis H1, i.e., that (past) changes in the intensity of dogwhistles in Flashback predict changes in the intensity of dogwhistles in Familjeliv.

Second, in support of H2, previous lags of changes in intensity in Familjeliv do not significantly predict changes in intensity in Flashback. This is observed for both the mean and the LOCF-imputed data (Table 2 and Table A1).

A third observation is that in Flashback and Familjeliv, there is a strong autocorrelation process, i.e. the present changes of intensity are significantly explained by previous changes of the same variable (again, this is observed for both the mean and the LOCF-imputed data).

Eq.	Var.	*berika*	*globalist*	*återvandring*
FB_t	FB_{t-1}	-0.676***	-0.521***	-0.614***
		(0.141)	(0.129)	(0.14)
	FB_{t-2}	-0.431**	-0.33*	-0.334*
		(0.166)	(0.13)	(0.158)
	FB_{t-3}	-0.286†		-0.225
		(0.147)		(0.147)
	FL_{t-1}	-0.007	-0.029	-0.02
		(0.025)	(0.023)	(0.04)
	FL_{t-2}	-0.012	0.003	-0.041
		(0.026)	(0.023)	(0.04)
	FL_{t-3}	-0.0		-0.013
		(0.027)		(0.041)
FL_t	FB_{t-1}	1.206	-1.505*	0.251
		(0.783)	(0.707)	(0.476)
	FB_{t-2}	1.218	-0.59	1.785***
		(0.921)	(0.714)	(0.537)
	FB_{t-3}	2.218**		1.355**
		(0.816)		(0.498)
	FL_{t-1}	-0.541***	-0.672***	-0.73***
		(0.138)	(0.127)	(0.137)
	FL_{t-2}	-0.575***	-0.495***	-0.574***
		(0.144)	(0.126)	(0.136)
	FL_{t-3}	-0.288*		-0.233†
		(0.147)		(0.138)
R^2 for FB_t		0.35	0.29	0.32
R^2 for FL_t		0.42	0.43	0.51

Table 2: Coefficients of VAR models (mean imputation of data). FB = Flashback, FL = Familjeliv. †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

The IRFs provide further, but partial, support for H1 and H2. IDF patterns for the mean and LOCF-imputed data are very similar (Figure 2 and Figure A2). For *berika* and *återvandring*, responses in Familjeliv from the shocks in Flashback are positive at lag 1 to 3. Following the positive response, we observe a negative response before the effect

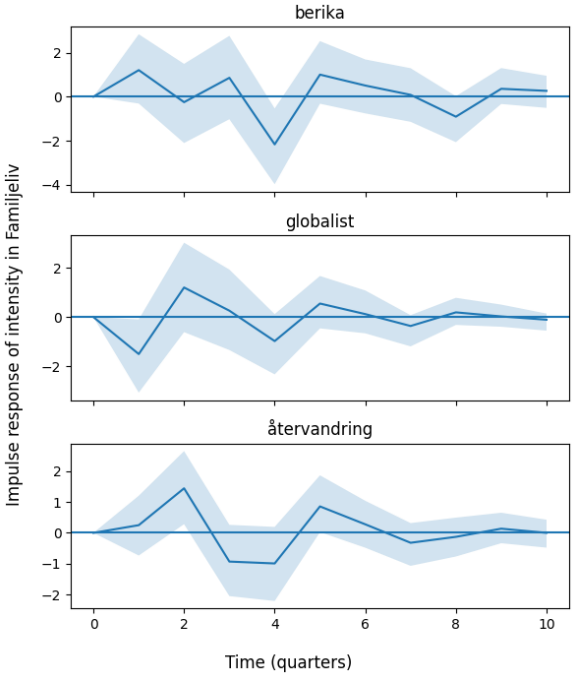

Figure 2: Impulse Response Functions (IRFs) from intensity in Flashback to intensity in Familjeliv (mean-imputed data). Filled area indicate standard errors at the 95% significance level.

fades away. That is, from a positive response to a shock, a negative response is predicted after the initial steps. As indicated already when discussing the coefficients for *globalist*, we at the first lag observe a negative response in Familjeliv to a shock in Flashback.

Since the time series of Flashback are not explained by the lagged values for Familjeliv (see Table 2), IRFs from Familjeliv to Flashback are for every DWE close to zero and not presented here.

We take the coefficients of the models and the IRFs to partially support H1 and H2. A test for Granger causality provides a stronger test of the temporal precedence of the intensity in Flashback relative to the intensity in Familjeliv.

In the cases of *berika* and *återvandring*, both with mean and LOCF imputation, changes in intensity in Flashback significantly Granger-cause changes in intensity in Familjeliv. For *berika*, mean imputed, $F(3, 90) = 2.84$, $p < 0.05$, and $F(3, 90) = 3.69$, $p < 0.05$, with LOCF. For *återvandring*, $F(3, 90) = 4.78$, $p < 0.01$, when mean imputed, and $F(4, 84) = 3.40$, $p < 0.05$, with LOCF. For *globalist* the null hypothesis of the Granger causality test (i.e., that changes in intensity in Flashback do *not* Granger cause changes in intensity in Familjeliv) cannot be rejected at the alpha-level of 0.05: $F(2,$

96) = 2.26, $p = 0.11$ (mean imputation) and $F(2, 96) = 0.57$, $p = 0.57$ (LOCF).

7 Discussion

Currently, there is only limited work that attempts to explain the changing patterns of dogwhistle behavior online (Åkerlund, 2022). Our results suggest that changes in dogwhistle behavior in communities with strong sympathies for the meaning encoded by the in-group meanings (in our case, racial bias and anti-immigration sentiment) predict the same in other less radical communities. More precisely, we find, first, that the intensity in Familjeliv can be predicted from the intensity three to nine months earlier in Flashback. As shown in the IRFs (Figure 2), a change in intensity in Flashback is followed by a response in Familjeliv in the following months. In the case of *berika* and *återvandring*, the test for Granger causality further confirms that changes in dogwhistle behavior in Flashback explain changes in dogwhistling in Familjeliv. Second, the opposite relationship does not hold. Changes in intensity in Flashback cannot be predicted from changes of intensity in Familjeliv. These observations support H1 and H2, although in the case of *globalist* we find only partial support. Flashback leads dogwhistle behavior online, while Familjeliv follows. This finding corroborates those in Åkerlund (2022). More generally, our findings support a process of mainstreaming far-right discourse (Klein, 2012; Munn, 2019; Mamié et al., 2021). Media such as Flashback may be acting as a gateway for adoption of anti-immigrant dogwhistling in the more mainstream communities such as Familjeliv.

The imputation of data used for *återvandring* and *globalist* is a concern for the robustness of the findings. However, the dynamics observed for *berika*, where there are only two zero-frequency quarters, supports our hypotheses. Although caution is required in interpreting the results for *återvandring* and *globalist* (due to imputation), they mostly point in the same direction as the more robust findings of *berika*.

One possible explanation for the difference between *globalist* and the other dogwhistles is that it is already associated with a long-standing anti-Semitic conspiracy theory associated with the financier George Soros (Langer, 2022). This is particularly prominent in online media. Members of Familjeliv may have already started to actively avoid it because of an association with anti-Semitism and conspiracy theory. Lindgren et al. (2024) picked it up as a dogwhistle because their population sample may include many members of the public less exposed to online political debate.

While we interpret the above findings in general framework of mainstreaming far-right discourse, much remains to be known about the underlying generative nature of this process. In the following, we raise two questions for discussion.

A first question is what explains the pattern of intensity in the more influential community, i.e. Flashback. In Flashback, temporal changes of dogwhistle usage are likely to be responses to prominent events, such as, elections, refugee flows, and the salience of political issues in mainstream media. Moreover, Flashback is part of a larger context of the dissemination of far-right content online (Klein, 2012), where it serves as a gateway between more radical racist forums and the mainstream media (Åkerlund, 2022). Thus, the salience of political events and issues can in turn be further amplified by "spin-doctors" (Sayeed et al., 2024) and intentional manipulation of digital discourses (Zhang et al., 2024). Moreover, internal dynamics of Flashback in terms of influential users is likely to drive the observed dogwhistle behavior (Åkerlund, 2021). The role of these factors (e.g., influential users, infiltration from extremist social movements, political events and the agenda setting of mainstream news media) and their interactions in explaining fluctuations of dogwhistle behavior in Flashback could be explored in future research.

Another question concerns the relationship between leaders and followers. What is the mechanism of diffusion? Previous work has found that radicalization of far-right ideas online occurs by internet users participating in incrementally more radical communities (Munn, 2019). For example, members of the so-called alt-lite movement consistently migrated to communities with more extreme views (Ribeiro et al., 2020). Participants of anti-feminist communities are likely to become members of far-right communities (Mamié et al., 2021). By observing and learning communication practices in a more radical community (Flashback), those behaviors can "brought back" to the origin community (Familjeliv). However, these processes are difficult to study directly. Although tracking user movement across communities *within* a platform is possible (but far from trivial), tracking user migrations *between* platforms (such as between

Familjeliv and Flashback) will not be reliable due to pseudonymity.

Limitations

A limitation of the present study is data scarcity. In the Swedish media environment, only a few dogwhistles are "active" for any given political moment. We used the dogwhistles from Lindgren et al. (2024) that were viable for analysis using the techniques we present in this paper. Producing more dogwhistles requires running a new selection process with no guarantee that a large "population" of dogwhistles can be found. Tracking a few instances of a phenomenon longitudinally is a standard approach in the study of political communication.

In Familjeliv, the DWEs are not used in all quarters analyzed. A natural response to this problem is to collect more data. However, this is not easily resolved when studying dogwhistle behavior, since dogwhistle usage is heavily skewed towards particular communities, with more limited usage across long periods of time in another community. Moreover, while the non-usage of DWEs are clear cases of zero frequency, how to interpret such cases in terms of intensity is less clear. Without an instance (word in context) there is no association to the in-group meaning to model. But these are not values missing (completely) at random. Given the fairly extensive need for imputation with *globalist* and *återvandring* in Familjeliv, the results of their models must be interpreted with caution.

In a sense, these are not "missing" values. Rather, the absence of a term is an observation in itself. The DWEs are much more common in Flashback than Familjeliv. With smaller proportions of the community sympathetic to the in-group, there is a lower probability of DWE usage. This is part of how dogwhistle behavior has evolved online.

Ethics Statement

Exploration of methodologies for the analysis of negative social phenomena always imply a risk that the tools developed will be used for malicious purposes, e.g. manipulating online political discourse. However, we believe that actors motivated to do so can do so anyway and that public research should not avoid the analysis of harmful communication for this reason. Rather, tools should be developed to detect, understand and explain these political behaviors to combat potentially harmful phenomena. Moreover, this work is part of the foundational work that is needed for understanding dogwhistle communication; it does not enable full detection on its own.

The corpus data used in this project were obtained from a national repository given responsibility for archiving Swedish documents of political and cultural significance. The replacement test survey was approved by the Swedish Ethical Review Authority.

Acknowledgements

Funding for this work was provided by the Gothenburg Research Initiative for Politically Emergent Systems (GRIPES) supported by the Marianne and Marcus Wallenberg Foundation grant 2019.0214 as well as a Swedish Research Council (VR) grant (2014-39) for the Centre for Linguistic Theory and Studies in Probability (CLASP). We wish to thank the anonymous reviewers for their constructive comments.

References

Mathilda Åkerlund. 2020. The importance of influential users in (re) producing swedish far-right discourse on twitter. *European Journal of Communication*, 35(6):613–628.

Mathilda Åkerlund. 2021. Influence Without Metrics: Analyzing the Impact of Far-Right Users in an Online Discussion Forum. *Social Media + Society*, 7(2):20563051211008831.

Mathilda Åkerlund. 2022. Dog whistling far-right code words: The case of 'culture enricher'on the Swedish web. *Information, Communication & Society*, 25(12):1808–1825.

Bethany L. Albertson. 2015. Dog-Whistle Politics: Multivocal Communication and Religious Appeals. *Political Behavior*, 37(1):3–26.

Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901.

Frank R Baumgartner and Bryan D Jones. 1993. *Agendas and instability in American politics*. University of Chicago Press.

Prashanth Bhat and Ofra Klein. 2020. Covert hate speech: White nationalists and dog whistle communication on twitter. *Twitter, the public sphere, and the chaos of online deliberation*, pages 151–172.

Helena Blomberg and Jonas Stier. 2019. Flashback as a rhetorical online battleground: Debating the (dis) guise of the Nordic Resistance Movement. *Social Media+ Society*, 5(1):2056305118823336.

Max Boholm, Björn Rönnerstrand, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Asad Sayeed. 2024. Can political dogwhistles be predicted by distributional methods for analysis of lexical semantic change? In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 144–157.

Max Boholm and Asad Sayeed. 2023. Political dogwhistles and community divergence in semantic change. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 53–65.

Janet M Box-Steffensmeier, John R Freeman, Matthew P Hitt, and Jon CW Pevehouse. 2014. *Time series analysis for the social sciences*. Cambridge University Press.

Ellen Breitholtz and Robin Cooper. 2021. Dogwhistles as inferences in interaction. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 40–46.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Herbert H. Clark. 1996. *Using Language*. Cambridge university press.

Katie Asplund Cohen, Björn Pelzer, Lisa Kaati, Nazar Akrami, Eric Andersson, and Felix Knutas. 2022. *En studie i fördom - Om rasistiska stereotyper i digitala miljöer*. Totalförsvarets forskningsinstitut, Stockholm.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert E Goodin and Michael Saward. 2005. Dog whistles and democratic mandates. *The Political Quarterly*, 76(4):471–476.

Linnea Hanell and Linus Salö. 2017. Nine months of entextualizations: Discourse and knowledge in an online discussion forum thread for expectant parents. In *Entangled Discourses: South-North Orders of Visibility*, pages 154–170. Routledge, New York.

Robert Henderson and Elin McCready. 2018. How dogwhistles work. In *New Frontiers in Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers 9*, pages 231–240. Springer.

Robert Henderson and Elin McCready. 2024. *Signaling without Saying: The Semantics and Pragmatics of Dogwhistles*. Number 17 in Oxford Studies in Semantics and Pragmatics. Oxford University Press, Oxford.

Niclas Hertzberg, Robin Cooper, Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, and Asad Sayeed. 2022. Distributional properties of political dogwhistle representations in Swedish BERT. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 170–175.

Giles Howdle. 2023. Microtargeting, dogwhistles, and deliberative democracy. *Topoi*, 42(2):445–458.

Jon Hurwitz and Mark Peffley. 2005. Playing the Race Card in the Post-Willie Horton Era: The Impact of Racialized Code Words on Support for Punitive Crime Policy. *The Public Opinion Quarterly*, 69(1):99–112.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Adam Klein. 2012. Slipping racism into the mainstream: A theory of information laundering. *Communication Theory*, 22(4):427–448.

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent signals, loud impact: Llms for word-sense disambiguation of coded dog whistles. *arXiv preprint arXiv:2406.06840*.

Armin Langer. 2022. Dog-whistle politics as a strategy of american nationalists and populists: George soros, the rothschilds, and other conspiracy theories. In Carsten Schapkow and Frank Jacob, editors, *Nationalism and Populism*, pages 157–187. Walter de Gruyter GmbH, Berlin.

Elina Lindgren, Björn Rönnerstrand, Gregor Rettenegger, Ellen Breitholtz, Robin Cooper, and Asad Sayeed. 2024. Coded appeals and political gains: Exploring the impact of racial dogwhistles on political support. *Journalism & Mass Communication Quarterly*, Epub ahead of print(0):10776990241280373.

Nicolás Lo Guercio and Ramiro Caso. 2022. An account of overt intentional dogwhistling. *Synthese*, 200(3):203.

Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2022. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7.

Helmut Lütkepohl. 2005. *New introduction to multiple time series analysis*. Springer.

Christoph Lutz and Christian Pieter Hoffmann. 2017. The dark side of online participation: exploring non-, passive and negative participation. *Information, Communication & Society*, 20(6):876–897.

Karl Malmqvist. 2015. Satire, racist humour and the power of (un) laughter: On the restrained nature of Swedish online racist discourse targeting EU-migrants begging for money. *Discourse & Society*, 26(6):733–753.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Robin Mamié, Manoel Horta Ribeiro, and Robert West. 2021. Are anti-feminist communities gateways to the far right? evidence from reddit and youtube. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 139–147.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. *Preprint*, arXiv:2305.17174.

Luke Munn. 2019. Alt-right pipeline: Individual journeys to extremism online. *First Monday*, 24(6).

Nicholas Niako, Jesus D. Melgarejo, Gladys E. Maestre, and Kristina P. Vatcheva. 2024. Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with arima and lstm. *BMC Medical Research Methodology*, 24:320.

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson. 2021. Semantic shift in social networks. In *Proceedings Of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37.

Thorsten Quandt. 2018. Dark participation. *Media and communication*, 6(4):36–48.

Anne Quaranto. 2022. Dog whistles, covertly coded speech, and the practices that enable them. *Synthese*, 200(4):330.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Faton Rekathati. 2021. The KBLab Blog: Introducing a Swedish Sentence Transformer.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141.

Patrícia Rossini. 2020. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49.

Asad Sayeed, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Björn Rönnerstrand. 2024. The utility of (political) dogwhistles–a life cycle perspective. *Journal of Language and Politics*.

Christopher A Sims. 1980. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. Language Science Press Berlin.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Rachel Wetts and Robb Willer. 2019. Who is called by the dog whistle? Experimental evidence that racial resentment and political ideology condition responses to racially encoded messages. *Socius*, 5:2378023119866268.

Ismail K White. 2007. When race matters and when it doesn't: Racial group differences in response to racial cues. *American Political Science Review*, 101(2):339–354.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Blow the dog whistle: A Chinese dataset for cant understanding with common sense and world knowledge. *arXiv preprint arXiv:2104.02704*.

Menghan Zhang, Xue Qi, Xinyan Liu, and Ke Zhang. 2024. Who leads? who follows? exploring agenda setting by media, social bots and public in the discussion of the 2022 south korean presidential election. *SAGE Open*, 14(2):21582440241248891.

A Appendix

A.1 Examples: high and low intensity

Warning: The following examples contain content that may be upsetting or offensive to some readers.

Examples are identified by measuring the intensity of individual sentences, i.e. the angular similarity of their sentence embedding and the

in-group embedding, as defined above. Examples are picked from the collection of the top 3 (high intensity) or bottom 3 (low intensity) of sentences in each quarter.

A.1.1 Enrich (*berika*)

Low intensity (= 0.50)

(1) Sedan tycker inte jag att det är fel att välja att vara hemma med barn, tror det kan **berika** livet mkt mer än hög lön i längden (Familjeliv, 2010, Q3)
(Then I don't think it's wrong to choose to stay at home with children, I think it can **enrich** life much more than a high salary in the long run)

High intensity (= 0.68)

(2) Ungdomsgäng som **berikar** invånarna med våld (Flashback, 2017, Q3)
(Youth gangs that **enrich** residents with violence)

A.1.2 Globalist (*globalist*)

Low intensity (= 0.58)

(3) Min far hade inte haft något emot ifall jag hade fostrats på ett indiskt sätt, då han är en **globalist** i den mening att han älskar mångkulturalism (Flashback, 2010, Q3)
(My father would not have minded if I had been raised in an Indian way, as he is a **globalist** in the sense that he loves multiculturalism)

High intensity (= 0.72)

(4) Juden skapar och finansierar blm rörelsen för att trycka tillbaka den vita rasen i sann **globalist** vänsteranda (Flashback, 2023, Q4)
(The Jew creates and finances the BLM [i.e. Black Lives Matter] movement to push back the white race in true **globalist** leftist spirit)

A.1.3 Re-migration (*återvandring*)

Low intensity (= 0.64)

(5) Sen får man ju ta med **återvandringen** i beräkningarna också, det är ganska många som faktiskt flyttar tillbaks igen om/när det lugnar ner sig i deras hemländer (Familjeliv, 2010, Q1)
(Then you have to include **re-migration** in the calculations as well, there are quite a few who actually move back again if/when things calm down in their home countries)

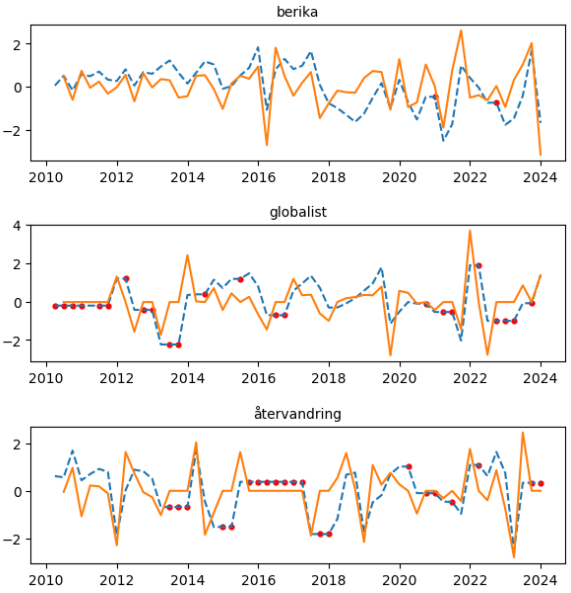

Figure A1: Intensity (dashed blue) and change of intensity (orange) and in Familjeliv (normalized by z-scores), with LOCF imputation of zero-frequency quarters (red dot).

High intensity (= 0.75)

(6) Den stora **återvandringen** ska påbörjas, samtliga som har på något sätt utnyttjat asylrätten ska deporteras tillbaka (Flashback, 2019, Q3)
(The great **re-migration** will begin, everyone who has in any way exercised the right to asylum will be deported back)

A.2 Visualization and results for LOCF transformation of data

Figure A1 shows the time series for the three DWEs with LOCF imputation. Table A1 shows the coefficients of the VAR models, with the LOCF-imputed data. Figure A2 shows the IRFs for the response in Familjeliv from a shock in Flashback.

Eq.	Var.	*berika*	*globalist*	*återvandring*
FB_t	FB_{t-1}	-0.676***	-0.529***	-0.55***
		(0.141)	(0.129)	(0.149)
	FB_{t-2}	-0.427**	-0.308*	-0.331*
		(0.167)	(0.129)	(0.168)
	FB_{t-3}	-0.28†		-0.254
		(0.148)		(0.183)
	FB_{t-4}			-0.03
				(0.159)
	FL_{t-1}	-0.013	-0.028	-0.013
		(0.025)	(0.021)	(0.038)
	FL_{t-2}	-0.013	0.012	-0.02
		(0.027)	(0.021)	(0.039)
	FL_{t-3}	-0.012		-0.009
		(0.027)		(0.036)
	FL_{t-4}			-0.004
				(0.038)
FL_t	FB_{t-1}	1.38†	-0.621	0.121
		(0.765)	(0.869)	(0.553)
	FB_{t-2}	1.367	0.365	1.54*
		(0.906)	(0.864)	(0.623)
	FB_{t-3}	2.465**		0.473
		(0.802)		(0.682)
	FB_{t-4}			-0.931
				(0.591)
	FL_{t-1}	-0.463***	-0.161	-0.358*
		(0.138)	(0.141)	(0.141)
	FL_{t-2}	-0.51***	-0.278*	-0.317*
		(0.145)	(0.142)	(0.144)
	FL_{t-3}	-0.248†		-0.235†
		(0.148)		(0.136)
	FL_{t-4}			-0.338*
				(0.141)
R^2 for FB_t		0.35	0.30	0.31
R^2 for FL_t		0.40	0.10	0.32

Table A1: Coefficients of VAR models (LOCF imputation of data). FB = Flashback, FL = Familjeliv. †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

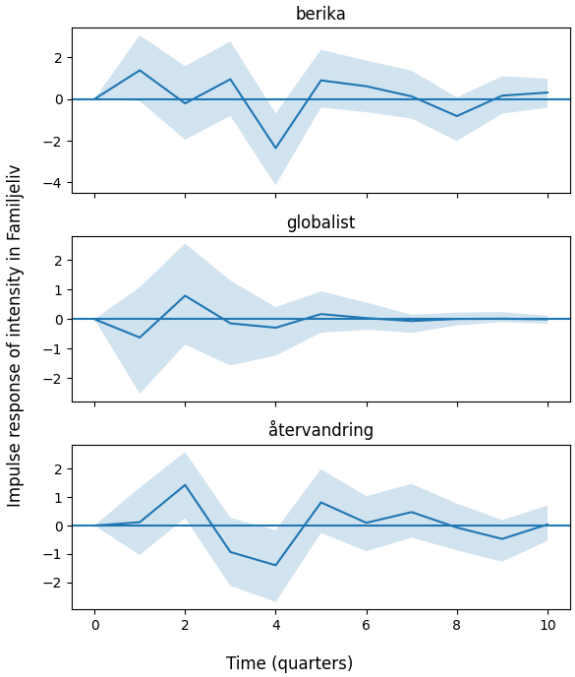

Figure A2: Impulse Response Functions (IRFs) from intensity in Flashback to intensity in Familjeliv (LOCF-imputed data). Filled area indicate standard errors at the 95% significance level.